Development of Human Genome Editing Tools for the Study of Genetic Variations and Gene Therapies

Dissertation advisor: Dr. George M. Church Author: Luhan Yang

Development of Human Genome Editing Tools for the Study of Genetic Variations and Gene Therapies

Abstract

The human genome encodes information that instructs human development, physiology, medicine, and evolution. Massive amount of genomic data has generated an ever-growing pool of hypothesis. Genome editing, broadly defined as targeted changes to the genome, posits to deliver the promise of genomic revolution to transform basic science and personalized medicine. This thesis aims to contribute to this scientific endeavor with a particular focus on the development of effective human genome engineering tools.

Chapter 1 introduces the key topics on genome editing, with an emphasis on its implications, current status, and potential applications.

Chapter 2 describes the generation of reTALEs, a simplified form of TALENs, and the assembly of a pipeline to scarlessly edit human stem cells. We demonstrate the utility of this pipeline by generating hiPSCs with mutations in HIV resistance genes within 3 weeks.

Chapter 3 describes the generation of a novel RNA-guided human genome editing tool. We reprogrammed a type II bacterial CRISPR system to function in a human context, and demonstrated an efficient & multiplexable version of this approach in multiple cell types including human iPSCs. We compared the efficiency and specificity of CRISPR with

TALE and designed a new strategy to mitigate the off-target issues associated with CRISPR.

To expand our genome editing toolbox, **Chapter 4** describes the assembly of novel chimeric deaminases that perform sequence-specific genome editing without generating DSBs and the need to simultaneously provide replacement (i.e., donor) DNA. Targeted deaminases are both efficient and specific in *Escherichia coli* and human cells, presenting an alternative platform that can eventually be used in multiplex genome editing.

Chapter 5 describes our effort in combining genetically engineered iPSCs with organ-on-chip models to investigate the cellular etiology of disease and to identify potential therapeutic targets. We generated isogenic iPSCs carrying a mutation identified in cardiomyopathy patients. Cardiomyocytes derived from engineered hiPSC recapitulated disease abnormalities and engineered "heart on chip" tissues contracted poorly. Replacement of the defective gene product corrected these abnormalities.

We finally conclude with remarks on the future prospects for genome editing to expand our understanding of fundamental biology and to enhance the wellness of human beings.

Table of Contents

Acknowledgements

I owe many thanks to the countless people who have helped me over my Ph.D journey.

I must first thank my thesis advisor Prof. George Church for the rich support that he has provided through the numerous discussions, encouragement, and resources that have enabled me to pursue this body of this work. George has inspired me to be the best scientist I can be not only by his enthusiasm and vision in science but also by his generosity and determination in helping others. Learning from George has been my greatest pleasure in my doctoral education.

I would also like to thank Prof. David Van Vector, the chair of BBS program, and Prof. Connie Cepko, the chair of Leder Human Biology and Translational Medicine, for providing me with the opportunity to pursue my degree at Harvard. Through its unique emphasis on medicine, my experiences with the Leder Program imprinted upon me a continuing desire to translate research into real clinical practices. I also thank my Ph.D program mentor, Prof. Elizabeth Engle, for her gracious support when I had just arrived at Harvard from China.

Many other great mentors have helped me grow over the past few years. My dissertation advisory committee members: Prof. Jon Seidman, Prof. Marc Vidal, and Prof. Ralph Scully have been extremely supportive in guiding my academic development and have provided critical guidance for my dissertation. I also thank Prof. Jon Seidman for his willingness to serve on my thesis defense committee. I thank Prof. Joseph Loscalzo and Prof. Anders Naar for allowing me to do research rotation in their labs and

explore my scientific interests. I need to thank my Peking University undergraduate mentor, Prof. Jingdong Zhao who had instilled in me the confidence and passion to pursue scientific career.

I am also grateful to have the opportunity to work with and learn from the phenomenal group of colleagues in the lab. I owe a big thanks to John Aach for helping and teaching me to analyze and present scientific results. I want to thank Prashant Mali for working closely with me on many faces of human genome editing projects and for building a role model for me as a creative and solid scientist, and to Marc Guell for inspiring me with his contagious enthusiasm in science and for his help in developing the human genome editing pipelines. I thank many people I have worked with in the Church laboratory so that I can fill up all the pages of this thesis: Adrian W Briggs, Susan Byrne, Joyce Yang, Xavier Rios, Po-Yi Huang, Raj Chari, Wei Leong Chew, Daniel Bryan Goodman, Venkataramanan Soundararajan, Caroline Kim-Kiselak, and David Cox. I also want to acknowledge these great people for the discussion and guidance that has helped me to grow professionally. I want to thank Billy Li for introducing me to this lab and challenging me to think independently; Srivatsan Raman for all his advices and encouragement that helped me persevere through difficult times; Sriram Kosuri for his witty suggestions. As a non-native English speaker, I also want to thank these people who have generously provided help on my writing: Bobby Dhadwar, Yoav Mayshar, Nikolai Eroshenko and Ben Stranges.

I also thank Prof. George Daley and Prof. William Pu, our wonderful collaborators at Harvard, for the many fruitful collaborations that bridge genome engineering with stem cell biology and disease modeling. I thank Alejandro De Los

Angeles from the Daley Lab, who has been extremely helpful for both discussion and experiments; Gang Wang from the Pu Lab for his many insights on using genome editing for drug screening and gene therapy. I thank John Hinson from the Seidmen Lab for introducing me a new world of cardiovascular biology. I also thank Prof. Keith Joung at MGH and Prof. Feng Zhang at MIT for providing advice that assisted the development of these tools.

I owe a big thanks to many great friends who helped me grow. I want to thank my closest friend Xuyu Cai and Jing Yang for putting up my ceaseless whining, laughing over my bad jokes and growing along with me. I thank Po-Yi Huang for giving me companionship during numerous late nights and weekends in the lab and Joyce for making the lab a truly fun place to work and play. I want to thank my hallmate Wataru Ebina for his aromatic Japanese cuisine and inspiring conversations. I also want to thank Rita Hu at Boston Consulting Group, who has made my short deviation from science a memorable and self-reflection journey.

Finally, I owe everything that I am to my parents, who bravely sent me out of my small town and support me to pursue my dream in a distant continent despite their loneliness and increasing frailty. I feel so blessed to have such loving parents while I feel guilty to not have too much time to be with them as the only child in the family. I only hope that their love and sacrifice have not gone in vain and I will live out of the dreams that their generation could not grasp.

Chapter 1

Introduction

"What I cannot create, I do not understand."

– Richard P. Feynman

Genome editing

Overview

A genome encodes instructions for executable actions that most organisms require to develop and respond to the environment. In 1928, a bacteriologist Frederic Griffith discovered that traits can be transferred between different strains of *Pneumococcus* by mixing dead bacteria with living recipients (*1*). The material that carries this information was suggested as DNA; one decade later Dr. Oswald Avery and his colleagues identified DNA as a "transforming principle" (*2*). Subsequently, DNA's role in the flow of genetic information was confirmed and stated in the central dogma of molecular biology (*3*) - "DNA makes RNA makes protein". Further technology advancements have enabled one to "read" the genome. More specifically, the development of Sanger sequencing in the 1970's (*4*) first allowed reliable analysis of DNA fragment sequences in the lab. Thereafter, the invention of DNA microarray and next-generation sequencing technology have continuously revolutionized the way genomic data is collected and perceived. Currently, the genomes of over 1000 organisms, including all three main domains of life (bacteria, archaea, and eukaryote) as well as many viruses, phages, plasmids and organelles have been sequenced (*5*). In addition, the number of annotations deposited in the Genebank database is growing at a breathtaking pace.

With the resulting wealth of information, scientists are poised to deliver upon the promises of the genomic revolution to translate basic science to personalized medicine. Central to this task, massive amounts of data must be converted into a functional and clinically relevant form. Genome editing, the effort to introduce targeted and defined chromosomal changes, can greatly facilitate our progress towards this direction.

First, biology questions can be answered directly and more simply using genome editing tools. For example, the large scale genome wide association studies (GWAS) and whole genome sequencing have identified thousands of genetic variations associated with diseases (*6*). While the function of each variant can be explored with reporter constructs in human or non-human cells, a direct path would be to edit a variant directly and precisely in the genomes of human cells to examine if this results in detectable phenotypes relative to unedited cells. This may be especially powerful for human induced pluripotent stem cells (hiPSCs), which can be differentiated into defined cell types to check for phenotypes specific to certain types (*7*).

Second, genome editing can be used to produce useful organisms and cure human diseases. Since DNA encodes the instructions used in the development and functionality of almost organisms, the implications of genome editing are far reaching. We can use such a tool to modify industrial and agricultural relevant organisms and confer them with desired traits. In addition, genome editing enables the treatment of genetic disorders by enabling permanent correction of specific mutations in DNA that cause an associated human disease.

Tools for genome editing

Overview

Considerable progress has been made to develop effective genome editing tools. Two categories of these tools are discussed, homologous recombination (HR)-mediated genome editing and non-HR-mediated genome editing, with emphasis on the former, which is also the approach most commonly used.

HR-mediated genome engineering

HR is a DNA repair mechanism highly conserved across all three domains of life to accurately repair harmful breaks that occur in the genome. In this naturally occurring genetic recombination event, nucleotide sequences are exchanged between two similar or identical molecules of DNA. HR is also used to produce new combinations of DNA sequence during meiosis.

Using the endogenous HR machinery, DNA information carried by exogenously introduced cloned DNA can be effectively introduced into the chromosome. This type of genetic manipulation was first demonstrated in yeast (*8*)and later exploited in other biological systems.

Specific genome editing techniques applied in creation of the mouse models was first reported in the late 1980's to generate knock-out and knock-in mice (*9*). In 1985, the first study of gene targeting in human cells was reported (*10*). Introduction of exogenous DNA carrying homology arms against the native β-globin locus successfully inserted non-human elements into the defined region with 10^{-3} frequency (*10*). These specific one-copy insertions were not accompanied by insertions of exogenous sequences at other off-target genomic sites. However, the broad utility of this HR approach was greatly limited by its low efficiency.

The effect of double-strand break (DSB) on homologous recombination emerged from the study of yeast model system in 1980s (*11*). DSBs stimulate the cellular DNA repair mechanisms, including the error-prone nonhomologous end joining (NHEJ) and HR. It was found that introduction of a DSB within or near the targeting site enhances the rate of HR 10-100 folds in the yeast (*11*). In the following decades, a number of labs demonstrated that a specific DSB in the genomic target created by the I-SceI homing endonuclease stimulated HR between the genomic target and transfected plasmid (‘‘gene targeting’’) by 1,000-fold (*12*). With optimization, gene targeting rates of 3%–5% can be obtained.

Inspired by the discovery of the effect of DSB, scientists proceeded to devise methods to introduce DSBs at a gene of interest in a targeted fashion. One solution is to engineer nucleases that enable sequence specific cuttings at the target site. To this end, engineered nucleases must have a combination of qualities: first, an engineered nuclease should be sufficiently adaptable; second, it needs to recognize sufficiently long target sequence that is unique in complicated eukaryotic genome. Chimeric nucleases, composed of programmable DNA-binding domains, such as engineered zinc fingers (ZF) and transcription activator-like effectors (TALE), fused to a nonspecific DNA cleavage module, possess these critical features.

The zinc finger domain is the most abundant DNA-binding motif in the human genome and was the first DNA binding domain used in chimeric nucleases (*13*). It consists of a ββα protein configuration, in which the α-helix binds to the major groove of DNA and recognizes 3 bp contiguous nucleotides. A tandem array of zinc finger domain can bind to longer DNA sequence, such that a 3 zinc finger array has 9bp DNA recognition specificity. Zinc finger nucleases, fusing of a zinc finger array with the nuclease domain derived from the type IIS restriction enzyme FokI, were first developed by Chandrasegaran and his colleagues (*14*). Fok1

must dimerize to cleave DNA, thus cleavage by Fok1 as part of a ZFN-based system requires two adjacent and independent binding events, enabling specific targeting of long and potentially unique recognition sites (2X9=18). The DNA-binding specificity of zinc finger domain has been extensively engineered. Methods such as "modular assembly" (*15*), combinatorial selection, and "OPEN (Oligomerized Pool Engineering)" (*16*) have collectively, yielded unique zinc finger domains with specificity for almost all the possible nucleotide triplets. Commercialized zinc fingers are also available through Sangoma. Engineered ZFNs have been used to conduct genome editing in C. elegance, drosophalia, zebrafish, mouse, rat, catfish, sea urchin, rabbit, pig and corn (*17*). In 2003, Porteus and Baltimore demonstrated that ZFNs could enhance gene targeting by several-thousand–fold in human stem cells (*18*). Subsequently, it has been demonstrated in human primary T cells that ZFNs can mediate genome targeting on IL2RG gene with efficiency up to 5% (*19*). These results opened up the possibility of using ZFNs to product human genome editing. However, the prohibitive price and long selection process made ZFNs not accessible to the broad scientific communities.

Discovery of the elegant correlations between protein sequences of transcription activator-like effectors (TALEs) with their DNA binding sequence expanded the option of engineering a programmable DNA-binding protein (*20*). TALE is a naturally occurring protein originally from *Xanthomonas* bacteria. It mimics the transcriptional regulatory factors to hijack the host expression system. TALE carries a central DNA-binding domain, consisting of a repeating chain of nearly identical 34-amino acid monomers. Each monomer recognizes a single DNA target base with four common monomer variants optimally binding one of the four DNA base pairs. Like zinc finger domains, these TALE repeat can be linked together to recognize a specific contiguous DNA sequence (*21*). Several protocols have been developed to assemble

customized TALE, including golden gate cloning (*22*), hierarchical ligation (*21*) and solid phase synthesis (*23*, *24*). Customized TALEs have become commercially available through Cellectis Bioresearch (Paris, France) and Life Technologies (Grand Island, NY). Engineered TALE, when fused to other user-specified domains, can address a vast range of proteins and other molecules to particular genomic locations both *in vivo and in vitro*. TALEN, with TALE fusing with the catalytic domain of Fok I, have been engineered to generate novel DNA nucleases (*25*). Similar to ZFN, dimerization of TALENs with designated orientation and spacer is required to function at defined region. As of writing this thesis, TALENs have been widely applied to many organisms and cell lines (*24*, *26*).

Apart from the engineered chimeric nucleases described above, the recently developed Clustered, regularly interspaced, short palindromic repeat (CRISPR)/ CRISPR associated system (Cas) present an alternative approach for introducing sequence specific cutting in the genome. CRSPR/Cas, the adaptive immune system found in bacteria and archaea, uses short RNA to direct degradation of foreign DNA. In type II CRISPR/Cas sytem, short CRISPR RNA (crRNA) anneals with trans-activating crRNAs (tracrRNAs) and direct DNA cleavage at crRNA matching site by Cas9 protein (*27*, *28*). A recent *in vitro* reconstitution of the *Streptococcus pyogenes* type II CRISPR system demonstrated that crRNA fused with tracrRNA by a linker is sufficient to direct Cas9 protein to sequence-specifically cleave target DNA sequences matching the crRNA (*29*). The fully defined nature of this two-component system suggested that this system can be grafted into mammalian cell setting and used for genome editing.

<u>Non-HR-mediated genome editing tools</u>

An alternate approach of HR for achieving targeted genomic editing is the use of site specific recombinases (SSRs). SSRs can be categorized into two distinct two families, tyrosine recombinases, including most commonly used Cre and FLP, and serine recombinases (*30*). Despite different origins, SSRs rearrange DNA sequence using similar mechanism: the SSR catalyzes cutting at the recognition sites and rejoining of DNA strands to which it binds. This reaction promotes defined deletion, integration, inversion and cassette exchange. Given the coupling of chromosomal cutting with genetic material exchange, SSR has advantage over nucleases used to promote HR due to its elimination of NHEJ product (*31*).

Progresses have been made to reengineer SSR to act on novel sequence, but further work is needed to fully program SSR specificity. One approach to reprogram SSR exploits the modular nature of some serine recombinases. The catalytic domain and DNA binding domain are separated in some serine recombinase, in which the catalytic domain interacts with the central 14bp residues and the DNA binding domains recognizing the remaining 14bp. Thus, substitution of the DNA binding domain with customized DNA binding protein would confer the chimeric enzyme new sequence specificity, as demonstrated by zinc finger recombinases (*32*) and TALE recombinases (*33*). Nevertheless, the 14 bp fixed sequence specificity carried by catalytic domain still limits the broad targeting capacity of engineered recombinases. Selection among SSR mutants has also been conducted to obtain SSR variants acting on different central sequences, but success was only obtained on sequences that closely resemble the enzyme's nature sites (*30*).

Single-stranded oligodeoxyribonucleotides (ssODNs) (*34*) are another alternative method for mammalian genome editing. The simple requirements of ssODNs for genome editing make ssODNs convenient to use and potentially highly multiplexible. ssODNs mediated genome

editing has proven very efficient to incorporate the information in *E.coli* genome via mimicking the okazaki fragment during DNA replication (*35*). The mechanism for ssODN-mediated gene targeting in mammalian cell is unclear. In addition, although mammalian cell system, site-specific sequence alterations have successfully been introduced into mouse embryonic stem cells (mESCs) and several transformed human cells using ssODNs (*36*), the targeting efficiency is still low, ranging from 10^{-3}~10^{-6} in a variety of human cells (*37*). The suppression of mismatch repair (MMR) activity can enable effective ssODN-mediated genome targeting in many cell types (*37*), but the level of undesired mutations accumulated in the modified cells during the brief period of MMR suppression is unknown. Future studies addressing the mechanism and enhancing the efficiency of ssODNs-mediated genome editing is ssODNs approach practical to use on the mammalian genomes.

Applications of genome editing

Overview

In addition to functional studies of the genome, the efficient alteration of genomic sequences in a wide range of organisms and cell types has inspired endeavors to use these tools in various academic, pharmaceutical and industrial settings. We will discuss the applications and prospects in 1) building model organisms, 2) agriculture and industrial productions and 3) therapeutic and pharmaceutical applications.

Applications in building animal/cellular models system

The ability to efficiently edit the genome has led to the development of new animal models and cellular systems for biological research and clinical applications. Genome manipulation has been conducted in commonly used model organisms (*38*) such as *Caenorhabditis elegans,* zebrafish, drosophila, rats and mice, in addition to more exotic species including frogs, sea urchins, cricket, rabbits and butterflies. Moreover, genome editing approaches have been extended to human pluripotent cell lines to model a broad range of genetic conditions (*39*). These model systems can help us study human diseases. For example, inactivation of the gene encoding low-density lipoprotein (LDL) receptor in pig models familial hypercholesterolemia (*40*). In addition, modified human stem cells carrying specific diseases mutations can be used in the *in vitro* drug screening and toxicological tests that otherwise not directly feasible in human subject (*7*).

Applications in agriculture

Genome editing tools can be used to generate new agriculturally- relevant plants and animal species. It has been demonstrated that introduction of specific mutations and transgenic insertions can confer herbicide resistants (*38, 41*) in corn, tobacco and *Arabidopsis thaliana*. In addition, genetically modified live stocks have also been produced (*42*). For example, 'enviropigs', produced by the addition of the Phytase gene of *E.coli* origin in pigs, is capable of digesting phytic acid in the cereal grains, thus reducing feed costs and phosphorus pollution.

Applications in gene therapy

The ability to edit the human genome presents opportunities for treatment of genetic disorders. In general, there are two different approaches, 1) stem cell based *ex-vivo* gene therapies and 2) direct *in vivo* gene therapies.

The idea of human stem cell-based gene therapy is centered around the prospect of restoring normal gene function under the control of endogenous regulatory elements and generating a ready supply of genome corrected cells for autologous transplantation (*43, 44*). Although the concept of stem cell based gene therapy seems futuristic, scientists have made considerable progress on various fronts. First, following the exciting study conducted by Yamanaka's group (*45*), several non-transgenic approaches have been developed to reprogram human somatic cells into human induced pluripotent stem cells (hiPSCs) (*46*). Second, genomic defects within hiPSCs can be corrected by the deployment of site-specific nucleases and DNA donors. It has been demonstrated that monogenicdisorders, such as sickle cell anemia (*47*), cystic fibrosis (*48*), Huntington's diseases(*49*), Parknison's (*50*), X- linked severe combined immune deficiency (SCID) (*51*) and hemophilia B (*52*) can be genetically corrected. Third, hiPSCs demonstrate the ability to differentiate into many defined somatic cell types of the body (*43*), including hepatic, pancreatic, intestinal, pulmonary, neural progenitor cells, haematopoietic cells and cardiomyocyte for autologous transplantation. In addition, the therapeutic value of gene editing for stem cell-based therapy has been demonstrated in mouse models. For example, Jaenisch and colleagues used homologous recombination to repair the genetic defect in iPS cells derived from a humanized mouse model of sickle-differentiation (*53*); subsequently, it was demonstrated that transplantation of the corrected haematopoietic progenitors into the affected mice rescued the disease phenotype.

In vivo genome editing is another promising approach for the treatment of genetic disorders. It has been demonstrated in a mouse model of haemophilia that ZFNs are able to induce DSBs efficiently when delivered directly to mouse liver (*52*). The level of gene targeting achieved was sufficient to correct the prolonged clotting times in mice, and remained persistent after induced liver regeneration.

Despite heady progress, to realize the potential of using gene therapy tools in clinical practice, several technique hurdles must be addressed. First, it is important to demonstrate that no oncogenic mutations have been made in the corrected cells or transfected tissues. Non-specific DSBs introduced by customized nucleases are known to be mutagenic and can lead to unspecific regional mutations and chromosomal translocations. Genome wide sequencing has been performed to investigate the off-target mutations generated by ZFNs and only a few mutations were identified in the colonized cells (*47*). However, this aspect is not clear for TALEN edited cells, or CRISPR targeted systems. It is critical, therefore, to systematically assess the specificity of the nucleases and engineer better system that minimizes these off-target effects. Second, for *in vivo* gene therapy, efficient and specific gene delivery methods as well as strategies to address potential adverse immune responses need to be investigated.

Aside from restoring the function of disease genes, genome editing tools can be used for other types of clinical applications. For example, genome targeting practices have been performed to disrupt C-C chemokine receptor type 5 (CCR5) (*54–56*), the HIV co-receptor, in hematopoietic stem cells, which confers differentiated T cells and transplanted mouse HIV resistance. This approach is currently in clinical trials (NCT01252641, NCT00842634 and NCT01044654). Besides that, the concept of epigenetic chromosome therapy has been successfully demonstrated. In this exciting study, insertion of a copy of XIST (the X-inactivation

gene) on the extra copy of chromosome 21 silent the host chromosome, thus correcting gene imbalance of pluripotent stem cells derived from Down's syndrome patients (*57*).

Development of new genome editing tools

Throughout this thesis, we described the development of novel genome editing tools: first, reTALEs (Chapter 2), which simplify tool synthesis and enable the elusive construction of lentivirus particles encoding TALE; CIRSPR (Chapter 3), a highly multiplexible and robust genome editing venue; and targeted demainases (Chapter 4), a genome editing tool that does not introduce DSBs and does not require DNA donors. Additionally, we investigated the genome editing specificity of TALE and CRISPR (Chapter 3) and devised strategies to mitigate off-target effects. Finally, we enhanced accessibility to these new methods by constructing a robust pipeline for scarless human stem cell genome engineering and demonstrated the utility of these tools by applying the engineered cell lines to study cellular etiology of human diseases.

References

1. F. Griffith, The significance of pneumococcal types, *Journal of Hygiene* **XXVII** (1928) (available at http://journals.cambridge.org/abstract_S0022172400031879).

2. M. MCCARTY, Studies on the chemical nature of the substance inducing transformation of pneumococcal types, *The Journal of Experimental Medicine* (1979) (available at http://130.14.81.99/ps/access/CCAAAM.pdf).

3. F. Crick, Central dogma of molecular biology, *Nature* (1970) (available at http://cs.brynmawr.edu/Courses/cs380/fall2012/CrickCentralDogma1970.pdf).

4. F. Sanger, S. Nicklen, DNA sequencing with chain-terminating, **74**, 5463–5467 (1977).

5. D. a Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, E. W. Sayers, GenBank., *Nucleic acids research* **38**, D46–51 (2010).

6. P. Marjoram, a Zubair, S. V Nuzhdin, Post-GWAS: where next? More samples, more SNPs or more biology?, *Heredity* , 1–10 (2013).

7. H. Zhu, M. W. Lensch, P. Cahan, G. Q. Daley, Investigating monogenic and complex diseases with pluripotent stem cells., *Nature reviews. Genetics* **12**, 266–75 (2011).

8. M. Shrivastav, L. P. De Haro, J. a Nickoloff, Regulation of DNA double-strand break repair pathway choice., *Cell research* **18**, 134–47 (2008).

9. A. Bradley, M. Evans, M. Kaufman, E. Robertson, Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines, *Nature* (1984) (available at http://www.nature.com/nature/journal/v309/n5965/abs/309255a0.html).

10. R. O. Smithies, M. Koralewski, Insertion of DNA sequences into the human chromosomal B-globin locus by homologous recombination, *Nature* **317**, 230 (1985).

11. a J. Klar, L. M. Miglio, Initiation of meiotic recombination by double-strand DNA breaks in S. pombe., *Cell* **46**, 725–31 (1986).

12. D. J. Segal, D. Carroll, Endonuclease-induced, targeted homologous extrachromosomal recombination in Xenopus oocytes., *Proceedings of the National Academy of Sciences of the United States of America* **92**, 806–10 (1995).

13. F. D. Urnov, E. J. Rebar, M. C. Holmes, H. S. Zhang, P. D. Gregory, Genome editing with engineered zinc finger nucleases., *Nature reviews. Genetics* **11**, 636–46 (2010).

14. Y. G. Kim, J. Cha, S. Chandrasegaran, Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain., *Proceedings of the National Academy of Sciences of the United States of America* **93**, 1156–60 (1996).

15. M. S. Bhakta *et al.*, Highly active zinc-finger nucleases by extended modular assembly., *Genome research* , 530–538 (2013).

16. M. L. Maeder *et al.*, Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification., *Molecular cell* **31**, 294–301 (2008).

17. P. Perez-Pinera, D. G. Ousterout, C. A. Gersbach, Advances in targeted genome editing., *Current opinion in chemical biology* **16**, 268–77 (2012).

18. M. H. Porteus, D. Baltimore, Chimeric nucleases stimulate gene targeting in human cells., *Science (New York, N.Y.)* **300**, 763 (2003).

19. F. D. Urnov *et al.*, Highly efficient endogenous human gene correction using designed zinc-finger nucleases., *Nature* **435**, 646–51 (2005).

20. J. Boch *et al.*, Breaking the code of DNA binding specificity of TAL-type III effectors., *Science (New York, N.Y.)* **326**, 1509–12 (2009).

21. F. Zhang *et al.*, LETTErs Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription, **29**, 149–154 (2011).

22. T. Cermak *et al.*, Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting., *Nucleic acids research* **39**, e82 (2011).

23. A. W. Briggs *et al.*, Iterative capped assembly: rapid and scalable synthesis of repeat-module DNA such as TAL effectors from individual monomers., *Nucleic acids research* , 1–10 (2012).

24. D. Reyon *et al.*, FLASH assembly of TALENs for high-throughput genome editing, *Nature Biotechnology* **30**, 460–465 (2012).

25. J. C. Miller *et al.*, A TALE nuclease architecture for efficient genome editing., *Nature biotechnology* **29**, 143–8 (2011).

26. C. Mussolino, T. Cathomen, TALE nucleases: tailored genome engineering made easy., *Current opinion in biotechnology* **23**, 644–50 (2012).

27. R. Sapranauskas *et al.*, The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli., *Nucleic acids research* **39**, 9275–82 (2011).

28. B. Wiedenheft, S. H. Sternberg, J. a Doudna, RNA-guided genetic silencing systems in bacteria and archaea., *Nature* **482**, 331–8 (2012).

29. M. Jinek *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity., *Science (New York, N.Y.)* **337**, 816–21 (2012).

30. W. R. a Brown, N. C. O. Lee, Z. Xu, M. C. M. Smith, Serine recombinases as tools for genome engineering., *Methods (San Diego, Calif.)* **53**, 372–9 (2011).

31. J. San Filippo, P. Sung, H. Klein, Mechanism of eukaryotic homologous recombination., *Annual review of biochemistry* **77**, 229–57 (2008).

32. C. a Gersbach, T. Gaj, R. M. Gordley, A. C. Mercer, C. F. Barbas, Targeted plasmid integration into the human genome by an engineered zinc-finger recombinase., *Nucleic acids research* **39**, 7868–78 (2011).

33. A. C. Mercer, T. Gaj, R. P. Fuller, C. F. Barbas, Chimeric TALE recombinases with programmable DNA sequence specificity., *Nucleic acids research* **40**, 11163–72 (2012).

34. F. Radecke *et al.*, Targeted chromosomal gene modification in human cells by single-stranded oligodeoxynucleotides in the presence of a DNA double-strand break., *Molecular therapy : the journal of the American Society of Gene Therapy* **14**, 798–808 (2006).

35. H. H. Wang *et al.*, Programming cells by multiplex genome engineering and accelerated evolution., *Nature* **460**, 894–8 (2009).

36. M. Aarts, H. te Riele, Progress and prospects: oligonucleotide-directed gene modification in mouse embryonic stem cells: a route to therapeutic application., *Gene therapy* **18**, 213–9 (2011).

37. X. Rios *et al.*, Stable gene targeting in human cells using single-strand oligonucleotides with modified bases., *PloS one* **7**, e36697 (2012).

38. J. K. Joung, J. D. Sander, TALENs: a widely applicable technology for targeted genome editing., *Nature reviews. Molecular cell biology* **14**, 49–55 (2013).

39. D. Hockemeyer *et al.*, Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases., *Nature biotechnology* **27**, 851–7 (2009).

40. J. Wei *et al.*, Characterization of a hypertriglyceridemic transgenic miniature pig model expressing human apolipoprotein CIII., *The FEBS journal* **279**, 91–9 (2012).

41. T. Gaj, C. a Gersbach, C. F. Barbas, ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering., *Trends in biotechnology* **31**, 397–405 (2013).

42. D. F. Carlson *et al.*, Efficient TALEN-mediated gene knockout in livestock, , 1–6 (2012).

43. D. a Robinton, G. Q. Daley, The promise of induced pluripotent stem cells in research and therapy., *Nature* **481**, 295–305 (2012).

44. A. B. C. Cherry, G. Q. Daley, Reprogrammed cells for disease modeling and regenerative medicine., *Annual review of medicine* **64**, 277–90 (2013).

45. K. Takahashi *et al.*, Induction of pluripotent stem cells from adult human fibroblasts by defined factors., *Cell* **131**, 861–72 (2007).

46. L. Warren *et al.*, Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA., *Cell stem cell* **7**, 618–30 (2010).

47. J. Zou, P. Mali, X. Huang, S. N. Dowey, L. Cheng, Site-specific gene correction of a point mutation in human iPS cells derived from an adult patient with sickle cell disease., *Blood* **118**, 4599–608 (2011).

48. K. a High, Update on progress and hurdles in novel genetic therapies for hemophilia., *Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program* , 466–72 (2007).

49. M. C. An *et al.*, Genetic correction of Huntington's disease phenotypes in induced pluripotent stem cells., *Cell stem cell* **11**, 253–63 (2012).

50. F. Soldner *et al.*, Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors., *Cell* **136**, 964–77 (2009).

51. T. Mashimo *et al.*, E. A. A. Nollen, Ed. Generation of knockout rats with X-linked severe combined immunodeficiency (X-SCID) using zinc-finger nucleases., *PloS one* **5**, e8870 (2010).

52. H. Li *et al.*, In vivo genome editing restores haemostasis in a mouse model of haemophilia., *Nature* **475**, 217–21 (2011).

53. J. Hanna *et al.*, Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin., *Science (New York, N.Y.)* **318**, 1920–3 (2007).

54. N. Holt *et al.*, Human hematopoietic stem/progenitor cells modified by zinc-finger nucleases targeted to CCR5 control HIV-1 in vivo., *Nature biotechnology* **28**, 839–47 (2010).

55. N. Holt *et al.*, Human hematopoietic stem/progenitor cells modified by zinc-finger nucleases targeted to CCR5 control HIV-1 in vivo., *Nature biotechnology* **28**, 839–47 (2010).

56. E. E. Perez *et al.*, Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases., *Nature biotechnology* **26**, 808–16 (2008).

57. J. Jiang *et al.*, Translating dosage compensation to trisomy 21, *Nature* (2013), doi:10.1038/nature12394.

Chapter 2

reTALE for scarless human stem cell genome editing

Luhan Yang[1,2], Marc Guell[1], Susan Byrne[1,6], Joyce Yang[1,2,6], Alejandro De Los Angeles[3,6], Prashant Mali[1], John Aach[1] , Caroline Kim-Kiselak[2], Adrian W Briggs[1], Xavier Rios[1], Po-Yi Huang[1,4], George Daley[3], and George Church[1,5] *

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA
[2]Biological and Biomedical Sciences Program, Harvard Medical School, Boston, Massachusetts, USA
[3]Children's Hospital, Boston, Massachusetts, USA
[4]Chemistry and Chemical Biology program, Harvard, Cambridge, Massachusetts, USA
[5]Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge,
[6]These authors contributed equally to this work

Acknowledgements

We thank all the Church lab members for suggestion and support; and Daniel Gibson (J. Craig Venter Institute) for providing advice on assembly reactions. This work is supported by National Human Genome Research Institute Center for Excellence in Genomics Science (P50 HG003170, G.M.C.).

Author contribution

L.Y., M.G. and G.M.C conceived the study jointly with S.B., J.Y., A.D.L., P.M., and J.A. designed and performed the experiments with assistance from C.K., A.W.B., X.R., P.H.; L.Y. and J.A. wrote the manuscripts with the help of all the other co-authors. G.D. supervised A.D.L; G.M.C. supervised all aspects of the study. In particular, I devised the idea, optimized the reTALE assembly reaction, compared the reTALE v.s. TALE efficiency, and worked with Marc Guell devised GEAS, and worked with Susan to genotype monoclonalized hiPSCs.

This chapter contains work from the manuscript "Optimization of scarless human stem cell genome editing", published in *Nucleic Acid Research*, July 31, 2013, doi:10.1093/nar/gkt555. The text and figures were modified to fit the format of this dissertation.

Summary

Precise genome editing in human induced pluripotent cells (hiPSCs) with targeted nucleases promises to advance biomedical research and personalized therapies. However, despite much progress, technical barriers in nuclease synthesis, accurate assessments of editing, and isolation of correctly edited cells continue to impede progress. To address these problems, we developed a robust pipeline for scarless genome modification of hiPSCs within three weeks that integrates: (1) TALEs that were recoded (re-TALEs) to eliminate DNA repeats, (2) rapid one-pot assembly of re-TALE nucleases (re-TALENs), (3) sensitive genome editing assessment, and (4) isolation of scarlessly edited hiPSCs without selection. Using our pipeline, we synthesized and tested 15 re- TALEN pairs targeting the *CCR5* locus and achieved targeted homology directed repair (HDR) rates in hiPSCs of 0.3~1.8% with single stranded DNA oligonucleotides (ssODNs). DNA repeat elimination enabled generation of functional re-TALE coding lentiviruses. We also documented effects on genome editing efficiency of target site chromatin state and quantified impacts of ssODN design parameters.

Introduction

Understanding the mechanisms by which genetic variation contributes to phenotype and disease is a central goal of human genetics and will be critical to the development of personalized medicine. Despite the rapidly-growing knowledge-base on human genetic variants associated with human disease (*1–3*), the functional significance of most of these variants is unknown. While the function of each variant can be explored with reporter constructs in human or non-human cells, a clearer path would be to edit a variant directly and precisely in the

genomes of human cells to examine if this results in detectable phenotypes relative to unedited cells. This may be especially powerful for human induced pluripotent cells (hiPSCs), which can be differentiated into defined cell types to check for phenotypes specific to certain types (*4*). Such an approach, ensuring permanent correction of specific mutations, also presents an opportunity in the treatment of genetic disorders (*5*), such as sickle cell anaemia (*6*), α1-antitrypsin deficiency (*7*), X-linked SCID (*8*) and p53-related cancers (*9*). In addition, precise genome editing in concert with the pluripotency of hiPSCs provides new venues for infectious disease treatment. For example, disruption of the *CCR5* locus, the gene encoding the HIV co-receptor, in human stem cells has made differentiated T cells resistant to HIV virus infection (*10*, *11*). Clinical trials are underway with this approach and current data have demonstrated improvement in several clinical parameters while being well tolerated.

Currently, nuclease-mediated genome editing provides the most efficient way to precisely edit cells, including human cells (*12–14*). By generating a double-stranded DNA (dsDNA) cut near the sequence to be edited, while providing homologous donor DNA containing the intended changes, cells can be induced to repair the cut with the donor and incorporate the desired sequence change with efficiencies as high as ~50% (Urnov et al., 2005; Cade et al., 2012), although generally less for hiPSCs (*18–20*). The most common approach in recent years to targeting dsDNA cuts has been to generate Zinc Finger Nucleases (ZFNs), with ZF domains programmed to bind to a target site being fused to FokI nuclease domains (*15*, *21*) to cut that site. Recently, however, Transcription Activator-Like Effectors (TALE) are being increasingly adopted in place of ZFs, since TALES not only have the advantage of a much simpler design (*22*), but when fused to FokI to produce TALE-Nucleases (TALENs), are proven to be more specific and less toxic than ZFNs (*23*, *24*), A number of methods have been published for

synthesizing TALENs increasingly rapidly and efficiently (*24–27*). However, TALENs target particular DNA sequences by means of an array of many (~12-20+) Repeat Variable Diresidue (RVD) domains (one RVD per target bp) that are extremely similar (*22*). This repeat structure complicates TALEN synthesis. Current methods circumvent this problem by iterative assembly that require building up the arrays from smaller pieces (*24*, *25*). It might be desirable therefore to eliminate the repeats at the DNA sequence level, which could eliminate the need for iterative methods and enable a faster, simpler, and less expensive one-pot synthesis of extended RVD arrays. Eliminating the RVD repeats left behind by current iterative methods could also address important *post-synthesis* problems that arise because the DNA repeats remain, such as the generation of high titers of lentivirus containing RVD arrays, which is critical for delivering the gene targeting tools into many cell types and animals (*28*).

Finally, it remains difficult to isolate scarlessly genome-edited hiPSCs. Owing to our limited knowledge of genome editing efficiencies in hiPSCs, most current methods introduce selectable or screenable markers along with the edit to enable isolation of correctly altered cells (*18*, *19*), leaving the host genome with contaminating non-human elements, or requiring additional steps with complex constructs and secondary selections (*7*) to remove the undesired elements. Recent genome-editing works using ssODNs in combination with ZFNs (*29*) obtained isogenic hiPSCs with precise genome editing without selection (*30*), raising the prospects of simplified scarless genome engineering. Nevertheless, the precise efficiency, scope, and biological mechanisms of ssODN mediated HDR (*31*, *32*) remain elusive. In addition, while correctly edited cells can be isolated by cloning out single cells and sequencing for the targeted changes if correct alteration frequencies are on the order of ~1%, this is challenging for hiPSCs which grow poorly as isolated cells deprived of cell-cell contacts (*20*). Thus, special methods

that assist clonal isolation of hiPSCs will improve both the therapeutic and experimental potential of hiPSCs.

Here we address all of these problems by presenting an integrated and extensible pipeline for conducting precise genome editing, including four novel components:

(1) A platform to design customized recoded TALEs (re-TALEs) that minimizes DNA repeats

(2) A robust protocol to assemble re-TALE transcription factors and re-TALE nucleases (re-TALENs) in a one-hour, one-pot reaction

(3) A high-throughput sequencing-based system to accurately and comprehensively assess both NHEJ and HDR genome editing efficiency, where HDR is assessed using ssODN donor DNA

(4) An efficient and rapid platform to conduct genotype screening of monoclonal hiPSCs

We demonstrate our pipeline by designing, generating, and testing 30 re-TALENs targeted to 15 distinct loci upstream of the *CCR5* gene, a therapeutic gene target of HIV treatment, in both K562 cells and hiPSCs, and then isolating monoclonal correctly edited hiPSCs. With the accuracy and sensitivity provided by our assessment system, we found more than half of the re-TALENs/ssODN achieving 0.3%-1.8% HDR and 0.3%-1.3% NHEJ efficiency in hiPSCs, and 3%-37% HDR and 3%-79% NHEJ in K562 cells, suggesting that re-TALENs in conjunction with ssODNs constitute a facile and broadly applicable genome editing tool. We show that eliminating the repeats in our recoded TALENs does not reduce the efficiency of genome editing compared with non-recoded TALENs, and that generating high titers of functional re-TALE lentiviruses is possible. Additionally, we found that there is no correlation

between genome editing efficiency and target site DNase I hypersensitivity (HS) but that there is an inverse correlation of HDR and nucleosome occupancy. We demonstrate that our pipeline can aid the improvement of genome editing by identifying optimal ssODN DNA donor design parameters. Lastly, by integrating our tools, we demonstrated a robust pipeline for obtaining scarless genome edited hiPSCs without selection within 3 weeks.

Our work presents an efficient and integrated toolkit for the design, synthesis and assessment of re-TALEs for genome editing in general and a robust pipeline for genome engineering of hiPSCs. The pipeline is also extensible in that tools generated in our system can be used with other targeted genome manipulations and vice versa. Our genome editing pipeline will provide researchers, clinicians and technologists alike with a flexible and powerful method for conducting genome editing in biomedical studies and, ultimately, clinical practice.

Results

Design of re-TALEs for genome editing

TALEs recognize target DNA sequences by means of arrays of tandem repeats of a 33-34 amino acid RVD domain with two variable positions, which consecutively recognize and bind individual base pairs of their target DNA sequence (*22*). However, this system leads to extensive repeats within TALE DNA coding sequence, which complicates the synthesis of DNA constructs expressing them. To overcome this problem, we devised an algorithm to radically redesign the sequence of TALE RVD arrays. The algorithm uses the redundancy of the genetic code to design a set of recoded TALE DNA sequences (*re-TALE*) which generates identical RVD amino acid sequences without the repetition of DNA sequences individually and as a group. (Figure 2_1 A) We also incorporated additional constraints to ensure that the re-TALEs present low potential for

5’ mRNA secondary structure, and an adequate codon adaptation index for human expression. Re-TALEs encoding 16 tandem DNA recognition monomers, plus the final half RVD repeat (16.5), are devoid of any 12bp repeats (Figure 2_1 A). Notably, this level of recoding is sufficient to allow PCR amplification of any specific monomer or sub-section from a full-length re-TALE construct (Figure 2_1 B). Our re-TALE design algorithm and code is made available to the public. This code can be used to eliminate DNA repeats in other arrayed repeated protein domains, including alternative TALE monomers or novel TALE frameworks such as the Goldy framework (*33*).

Robust re-TALE-N/TALE-TF Assembly

The improved design of re-TALEs makes it possible to order them from gene synthesis companies using standard technology (*34*), without incurring the added costs or extra procedures that come with sequences containing many repeats. For investigators wishing to synthesize re-TALEs in-house, however, we developed a library of RVD dimer blocks and backbone constructs (Figure 2_2 A) for a robust and cost-effective

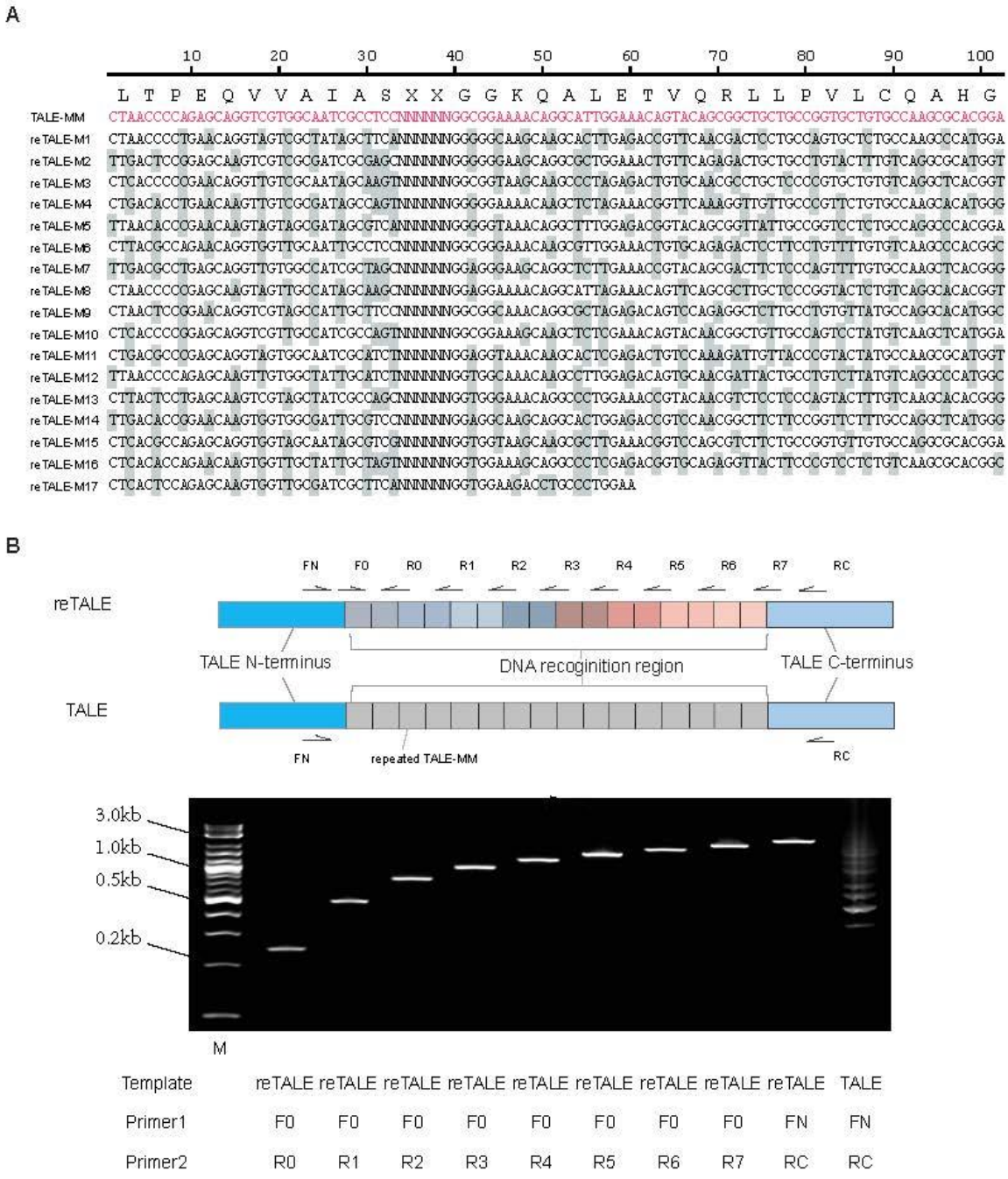


Figure 2_1. Design of reTALE
(A) Sequence alignment of the original TALE RVD monomer with monomers in re-TALE-16.5 (re-TALE-M1→re-TALE-M17). Nucleotide alterations from the original sequence are highlighted in gray.
(B) Test of repetitiveness of re-TALE by PCR. Top panel illustrates the structure of re-TALE/TALE and positions of the primers in the PCR reaction. Bottom panel illustrates PCR bands with condition indicated below. Note the PCR laddering presents with the original TALE template (right lane).

assembly protocol we call TASA (re-TALE Single-incubation Assembly). To enable specific assembly of re-TALEs, we designed the blocks such that each one shares unique 32 bp overlaps with adjacent blocks or the vector backbones, similar to the design of other assembly methods such as SLIC/ Gibson and CPEC (*35–37*). Unlike other TALE assembly methods, our re-TALE dimer blocks allow the full length re-TALEs to be efficiently assembled and amplified by PCR. Additionally, to simplify the procedure, we modified the destination vectors by incorporating *ccdB,* a bacterial negative selection cassette, flanked with paired endonuclease cutting sites at designated re-TALE cloning positions. With the concerted action of endonucleases, exonucleases, polymerases and ligases in the TASA reaction, re-TALE-TF/N plasmids can be directly assembled from re-TALE blocks and destination vectors. (Figure 2_2B) To reduce the frequency of false ligation products, we evaluated the activity of exonucleases at different enzyme concentrations (Figure 2_3) and chose Tth-derived ligase to increase ligation specificity (Figure 2_4). With optimized conditions, we assembled re-TALEs possessing 12.5, 14.5, 16.5 RVD monomers and assessed the assembly efficiency by checking the length of the cloned re-TALE insertions. We found perfect re-TALE assemblies with the following success rates: re-TALE-12.5, 46%; re-TALE-14.5, 32%; and re-TALE16.5, 18% (Figure 2_5). Detailed procedures for our robust and rapid TASA protocol can be found in the Methods. Sequences of the re-TALEs and backbone vectors are listed in the sequence Information, and all cloned re-TALE blocks and backbone vectors will be made publically available and will be deposited in Addgene.

<u>Comparison of re-TALEs and TALEs</u>

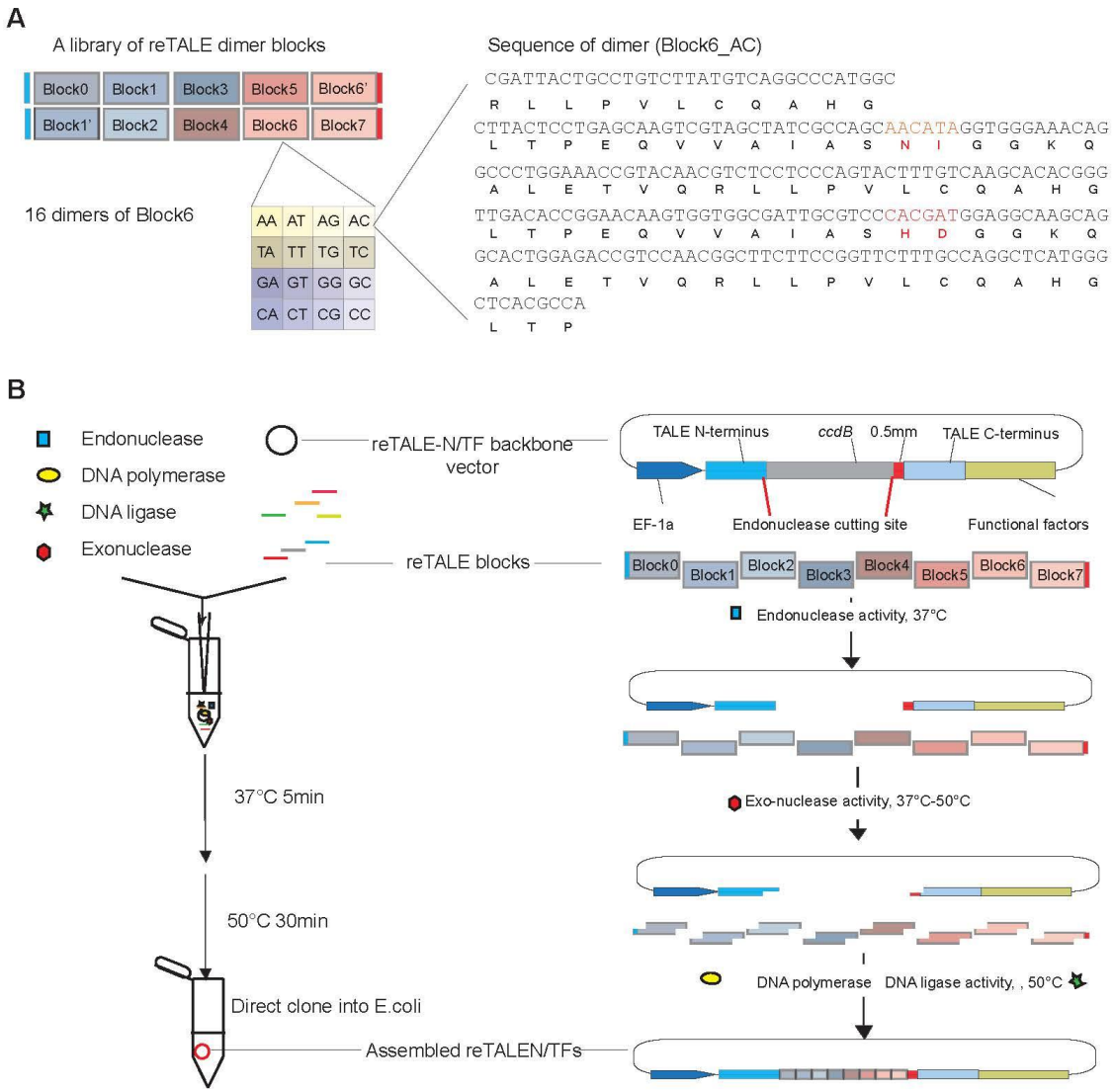


Figure 2_2. Design and practice of TASA assembly

(A) Schematic representation of the library of re-TALE dimer blocks for TASA assembly. There is a library of 10 re-TALE dimer blocks encoding two RVDs. Within each block, all 16 dimers share the same DNA sequence except the RVD encoding sequences; Dimers in different blocks have distinct sequences but are designed such that they share 32bp overlaps with the adjacent blocks. DNA and amino acid sequence of one dimer (Block6_AC) are listed on the right.

(B) Schematic representation of TASA assembly. The left panel illustrates the TASA assembly method: a one-pot incubation reaction is conducted with an enzyme mixture/re-TALE blocks/re-TALE-N/TF backbone vectors. The reaction product can be used directly for bacterial transformation. The right panel illustrates the mechanism of TASA. The destination vector is linearized by an endonuclease at 37° C to cut off *ccdB* counter-selection cassette; the exonuclease, which processes the end of blocks and linearized vectors, exposes ssDNA overhangs at the end of fragments to allow blocks and vector backbones to anneal in a designated order. When the temperature rises up to 50 ° C, polymerases and ligases work together to seal the gap, producing the final constructs ready for transformation.

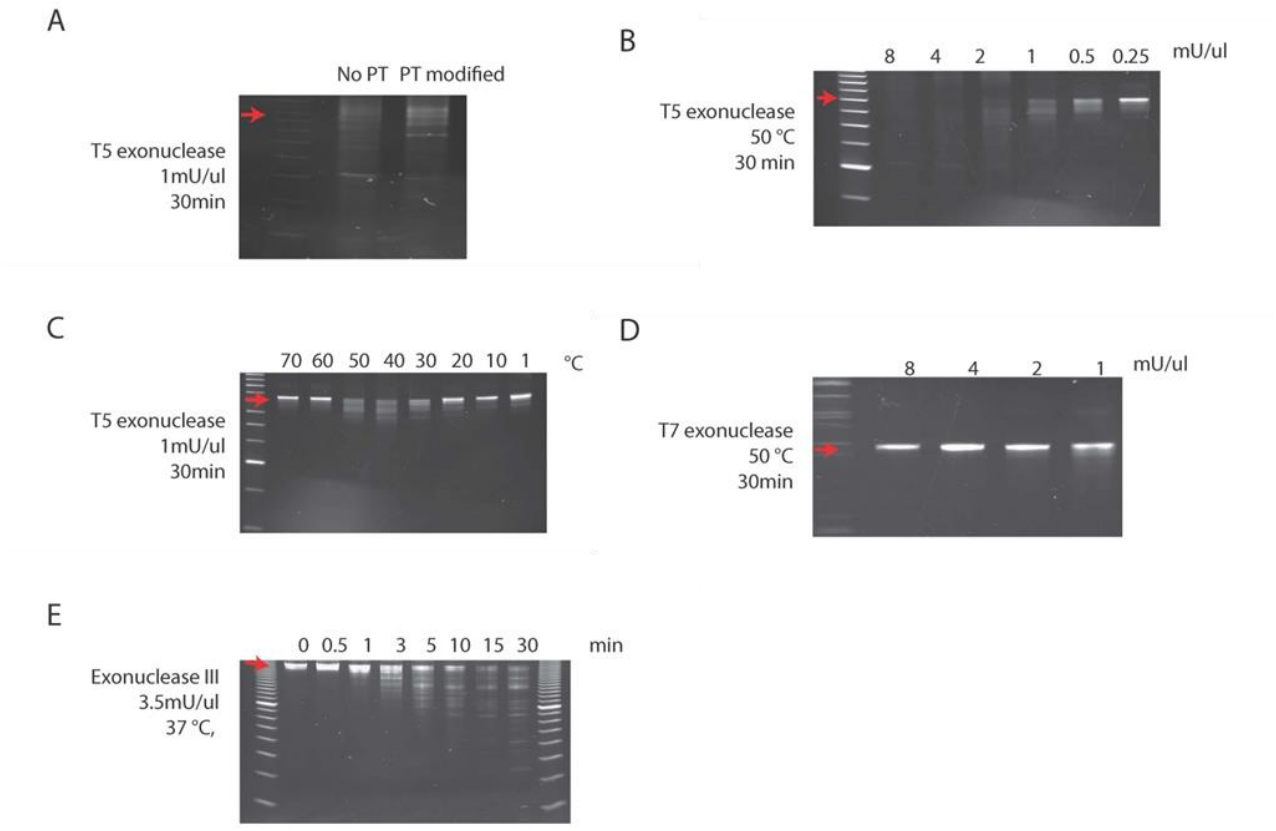


Figure 2_3. Optimization of exonuclease in TASA reaction

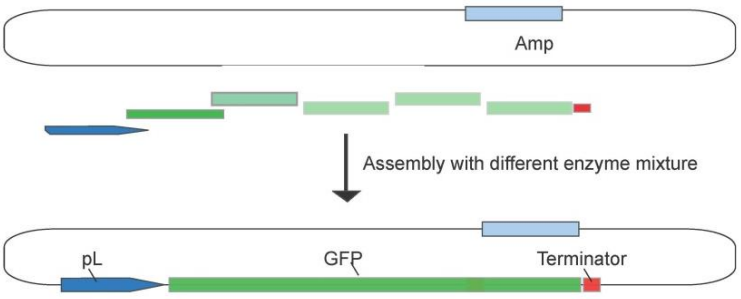
(A) To enable specific assembly, we sought to control the processivity of exonucleases. First, we tested whether phosphothioester (PT) linkage is resistant to T5 exonuclease activity so that we can use PT to stop the activity of exonuclease at designated sites. To this end, we embedded 3 consecutive PT bonds 30 nucleotides away from the ends of re-TALE blocks via PCR with PT modified primers and then incubated the re-TALE blocks with T5 exonuclease (T5 Exo) (1 mU/ µl). The size of digested ssDNAs was tested by running the reaction product on Urea-PAGE gels. We found that PT linkage is not resistant to T5 Exo-. Red-arrow indicates the position of full length re-TALE blocks.

(B) We next tested whether we can control the size of ssDNA overhangs by titrating the concentration of T5 Exo and we found 1mU/µl T5 Exo generates DNA overhangs around the size of 20-30NT at 50°C for 30min. We chose that condition thereafter.

(C) We then tested whether we can increase the reaction temperature to enhance the specificity of downstream DNA annealing and ligation while not compromising the T5 Exo activity. To this end, we tested the activity of T5 at different temperature and we found it is active within 30°C- 50°C, so we adhere to 50°C as the reaction condition.

(D) We then tested the activity of T7 exonuclease at 50°C and did not see any activity under such a condition.

(E) We tested the activity of Exonuclease III with different incubation times and we chose 1min as the reaction condition as Exonuclease III digests 10-30 nucleotides under such condition.



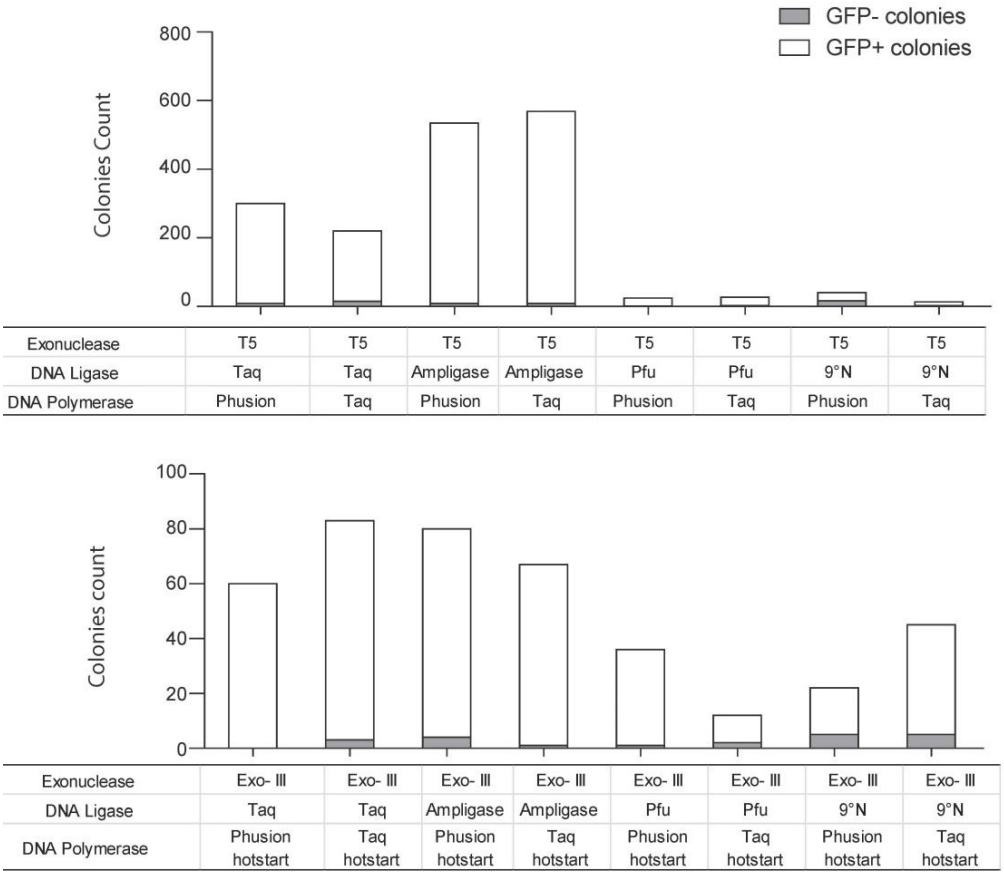


Figure 2_4. Optimization of ligases in TASA reaction

(A) Schematic representation of the experimental design for Optimization of ligases and polymerases in the TASA reaction. The diagram illustrates a GFP reporter which can be constructed with 6 pieces of DNA fragments and a destination vector using assembly reactions. We utilized the quick readout of colony numbers and GFP+ colony percentages of this reporter system to optimize the ligases and polymerases for the TASA reaction. pL: phage λ PL promoter.

(B) Test of efficiency and specificity of different assembly reactions. We tested different assembly enzyme mixtures (below), transformed enzyme mixture to E.coli and calculated the GFP+/GFP- colonies one day after transformation (above). We found that reaction containing Ampligase and Taq DNA polymerase yielded the most GFP+ colonies, suggesting the high efficiency and specificity of this enzyme mixture. The detailed enzyme protocol can be found in the Methods.

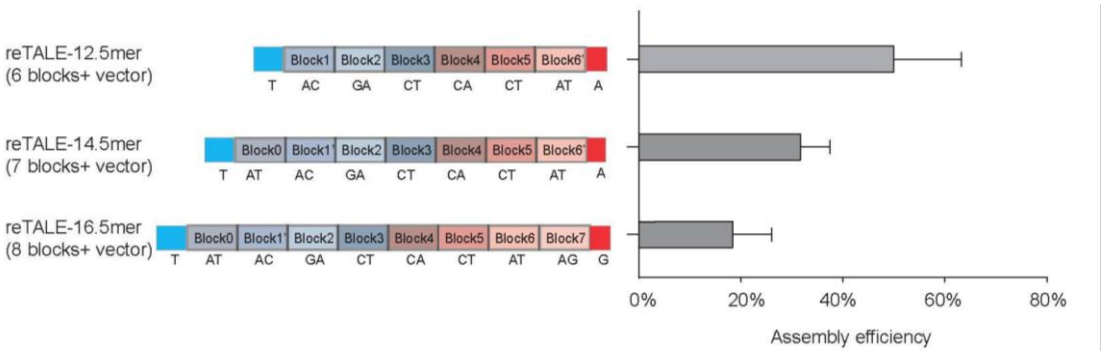


Figure 2_5. TASA assembly efficiency
TASA assembly efficiency for re-TALEs possessing different monomer lengths. The blocks used for assembly are illustrated on the left and the assembly efficiency is presented on the right.
equivalent levels of protein expression and activity. A re-TALE-TF (14.5mer) and re-TALE-TF

We next compared the activities of re-TALE-TFs and re-TALE-Ns against TALEs generated with non-recoded sequence in human cells. First, we constructed re-TALE-TF-2A-GFP and TALE-TF-2A-GFP plasmids coding for identical amino acid sequences, but where the latter was generated with non-recoded TALE components (*26*). These plasmids were transfected along with a mCherry reporter into 293T cells (Figure 2_6 A). We did not observe a significant difference in either GFP or mCherry expression between the two samples (Figure 2_6 C), indicating ses in (16.5mer) generated with longer sequence

recognition arrays to the same DNA target demonstrated similar levels of protein expression and transcriptional activation activity (Figure 2_6 B). DNA target demonstrated similar levels of protein expression and transcriptional activation activity (Figure 2_6 B).

Similarly, we constructed a pair of re-TALENs targeting the *PPP1R12C* (AAVS1) gene with the same amino acid sequences as previously published TALENs (*19*). To quantify the gene targeting efficiency, we built a 293T reporter cell line in which a chromosomally integrated mutant GFP gene can be repaired by nuclease-mediated HDR (*20*) (Figure 2_7 A) using this re-TALEN pair. re-TALEN-mediated HDR, as indicated by the percentage of GFP+ cells, exhibited an efficiency of 1.4%, similar to that of non-recoded TALENs (1.2%) (Figure 2_7 B). We next sought to verify the activity of re-TALENs on a native locus. To this end, we transfected PGP1 hiPSCs and 293T cells with the AAVS1 re-TALEN expression plasmids described above and a donor plasmid containing puromycin resistance and EGFP gene flanked by homologous sequence to the endogenous *PPP1R12C* gene (*19*) (Figure 2_7 C). We successfully obtained hiPSCs

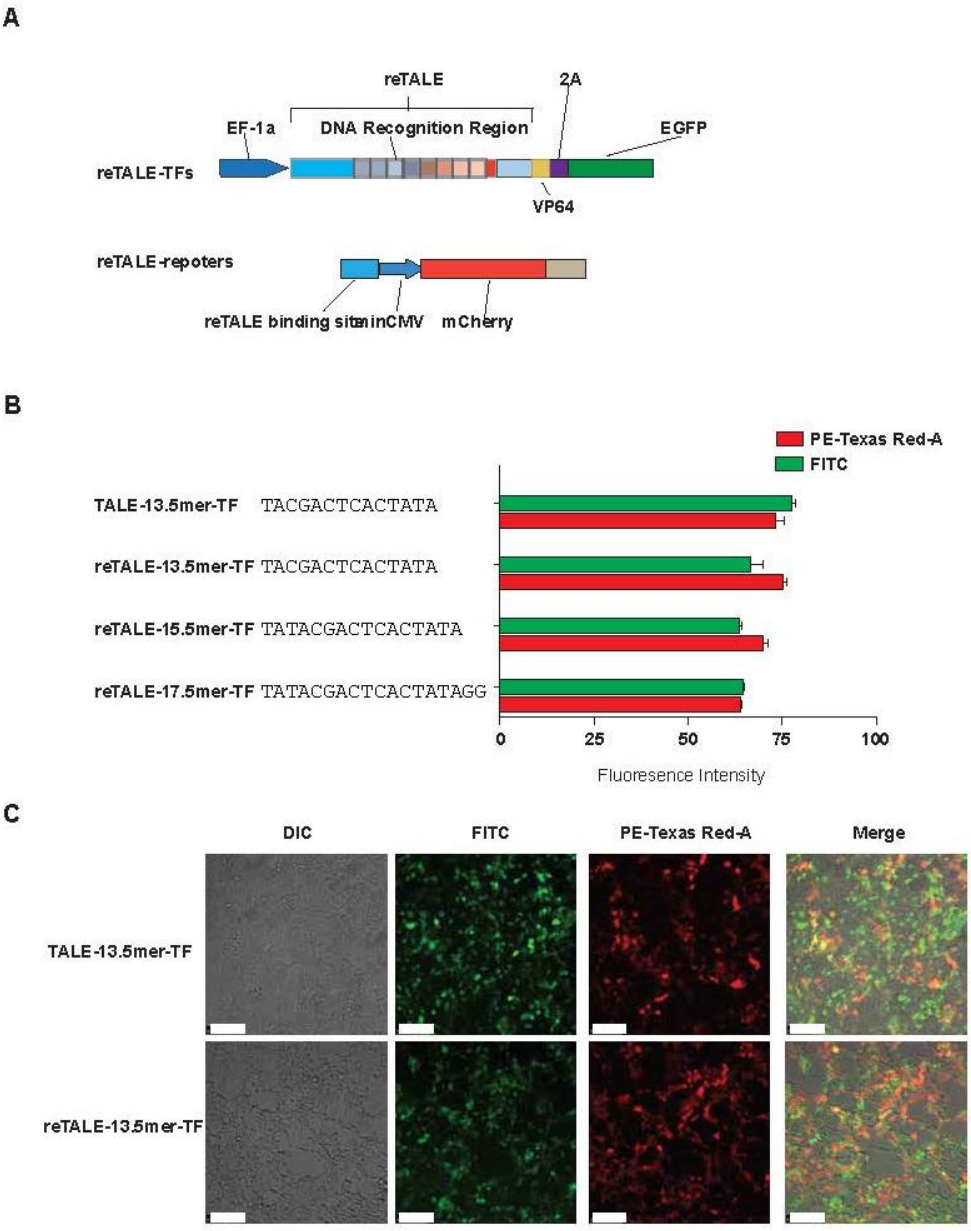


Figure 2_6. Expression level and activity of re-TALE-TFs

(A) Schematic representation of the fluorescence reporter system for testing re-TALE expression and activity. The diagram illustrates the structure of TALE- and re-TALE-TF-2A-GFP constructs and their mCherry reporters. VP64, synthetic transcription activation domain; 2A, self-cleavage peptides. Corresponding TALE-TF-2A-GFP constructs based on non-recoded TALE sequences were also constructed that coded for identical protein sequences.

(B) Expression level and activity of re-TALE-TFs: the binding site sequences of re-TALEs are shown on the left. The GFP and mCherry reporter expression levels of the corresponding re-TALE-TF-2A-GFP constructs were measured by flow cytometry (Methods). Fluorescent signal intensities are presented in arbitrary fluorescence units.

(C) Cells co-transfected with re-TALE-TF-2A-GFP plasmids and the reporter plasmid showed equivalent GFP and mCherry expression compared with the TALE-TF containing the same amino acid sequence. Scale bar, 100 μm.

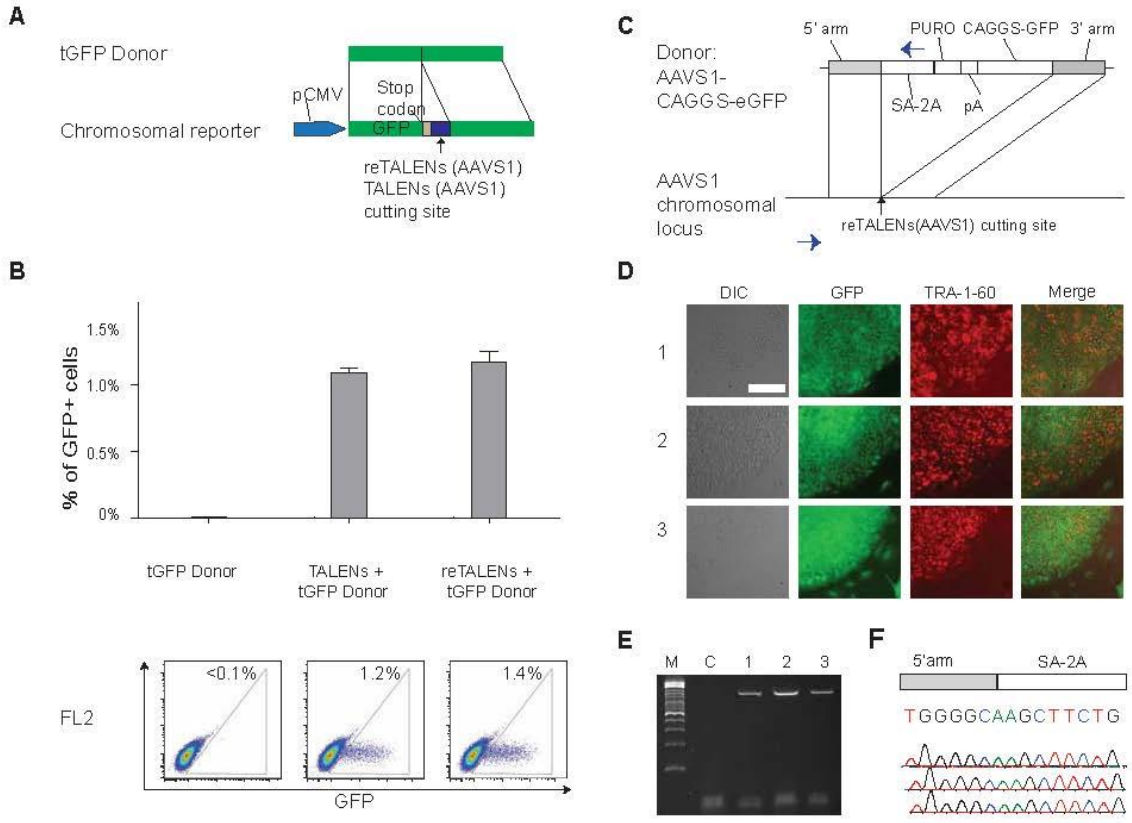


Figure 2_7. Efficiency of re-TALENs

(A) Schematic representation of experimental design for testing genome targeting efficiency. A genomically integrated GFP coding sequence is disrupted by the insertion of a stop codon and a 68bp genomic fragment derived from the AAVS1 locus (bottom). Restoration of the GFP sequence by nuclease-mediated homologous recombination with tGFP donor (top) results in GFP+ cells that can be quantitated by FACS. Re-TALENs and TALENs target identical sequences within AAVS1 fragments.

(B) Bar graph depicting GFP+ cell percentage introduced by tGFP donor alone, TALENs with tGFP donor, and re-TALENs with tGFP donor at the target locus, as measured by FACS. (N=3, error bar =SD) Representative FACS plots are shown below.

(C) Schematic overview depicting the targeting strategy for the native AAVS1 locus. The donor plasmid, containing splicing acceptor (SA)- 2A (self-cleaving peptides), puromycin resistant gene (PURO) and GFP were described before (Hockemeyer et al., 2011). The location of PCR primers used to detect successful editing events is depicted as blue arrows.

(Figure 2_7 D) from the pool of transfected cells after one week of puromycin selection and verified the specific genomic integration in hiPSCs by PCR and Sanger sequencing (Figure 2_7 E, F). Taken together, we concluded that re-TALE-N/TFs, despite their recoded DNA sequence, can effectively introduce genomic modifications in both somatic and pluripotent cells, at levels equivalent to TALENs and TALE-TFs with non-recoded sequences.

Next, we hypothesized that the removal of repeat sequences in re-TALEs would be beneficial for virus production. Lentiviral particles are powerful gene delivery vehicles for many cell types and *in vivo* animal studies (*38*, *39*). However, no study to date has reported generation of lentivirus carrying TALEs, probably due to the difficulty of generating functional viral particles encoding the repetitive TALE sequences. To test whether re-TALEs can improve the generation of functional lentivirus, we packaged lentiviral particles encoding re-TALE-2A-GFP or TALE-2A-GFP, and measured the viral titer based on GFP fluorescence. Re-TALE-TF-2A-GFP produced viral particles with titer of $1.4*10^6$ IFU/ml, 350X more than that of TALE-TF-2A-GFP ($4*10^3$ IFU) (Figure 2_8). To test the activity of re-TALE-TF encoded by viral particles, we transduced 293T cells with re-TALE-TF-GFP viral particles and transfected a mCherry reporter to the transduced 293T cell line 3 days after transduction. 293T cells transduced by lenti-re-TALE-TF showed considerably greater mCherry expression activation (Figure 2_8) compared with lenti-TALE-TF with equivalent titration.

Genome Editing Assessment System (GEAS)

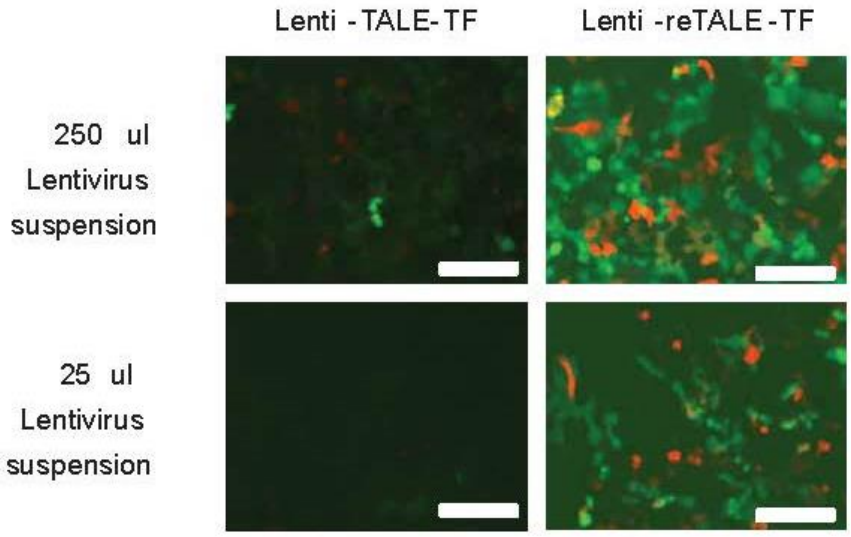


Figure 2_8. reTALEN activity and lentivirus generation potential
Images of lentivirus-transduced 293T cells transfected with mCherry reporter plasmid
We transduced 293T cells with lentiviral particles encoding re-TALE-TF-2A-GFP or TALE-TF-2A-GFP using 250 or 25 µl of lentiviral suspension. Three days after transduction, we transfected the transduced cells with the corresponding mCherry reporter to verify the activity of lenti-TALE-TF/lenti-re-TALE-TF. Scale bar, 100 µm.

Having validated the activity of re-TALENs, we next sought to build a sensitive and quantitatively accurate platform for simultaneously assessing re-TALEN-mediated NHEJ and HDR gene editing efficiencies, which we call GEAS (Genome Editing Assessment System). The idea is to deliver a pair of re-TALENs and an ssODN that matches the region of the re-TALEN target site except for a central 2bp mismatch, to allow re-TALEN dsDNA cutting and genomic repair to proceed, and then to conduct paired-end deep sequencing on the genomic region containing the target. HDR efficiency was measured by the percentage of reads containing exactly and only the 2bp mismatch within a 12bp window center of the re-TALENs target site. NHEJ efficiency was measured by the percentage of reads carrying indels. We developed a bioinformatics package in R to perform this analysis. To assess this system, we designed and constructed a pair of re-TALENs targeting the upstream region of CCR5 (re-TALEN pair #9 in Table S3) and a 90nt ssODN donor according to the specifications above (Figure 2_9 A), we then delivered the re-TALENs and ssODN into hiPSCs and K562 cells.

Delivery of ssODN alone into hiPSCs resulted in minimal HDR and NHEJ rates while the combination of re-TALENs with ssODN induced HDR with a rate of 0.67% and NHEJ with a rate of 0.73% (Figure 2_9 B). Gene editing is much more efficient in K562 cells, where we observed a 15% HDR rate and a 12% NHEJ rate, ~100X higher than the ssODN-only group (Figure 2_9 B). Notably, we observed that in both cell lines, the rate of genomic deletions and insertions, products of NHEJ, peaked in the middle of the spacer region between the two TALEN monomer sites for each of our re-TALENs (Figure 2_9 B), as would be expected from the fact that the DSB takes place in this region. We observed a median deletion size of 4bp and insertion of 3bp in hiPSCs and a

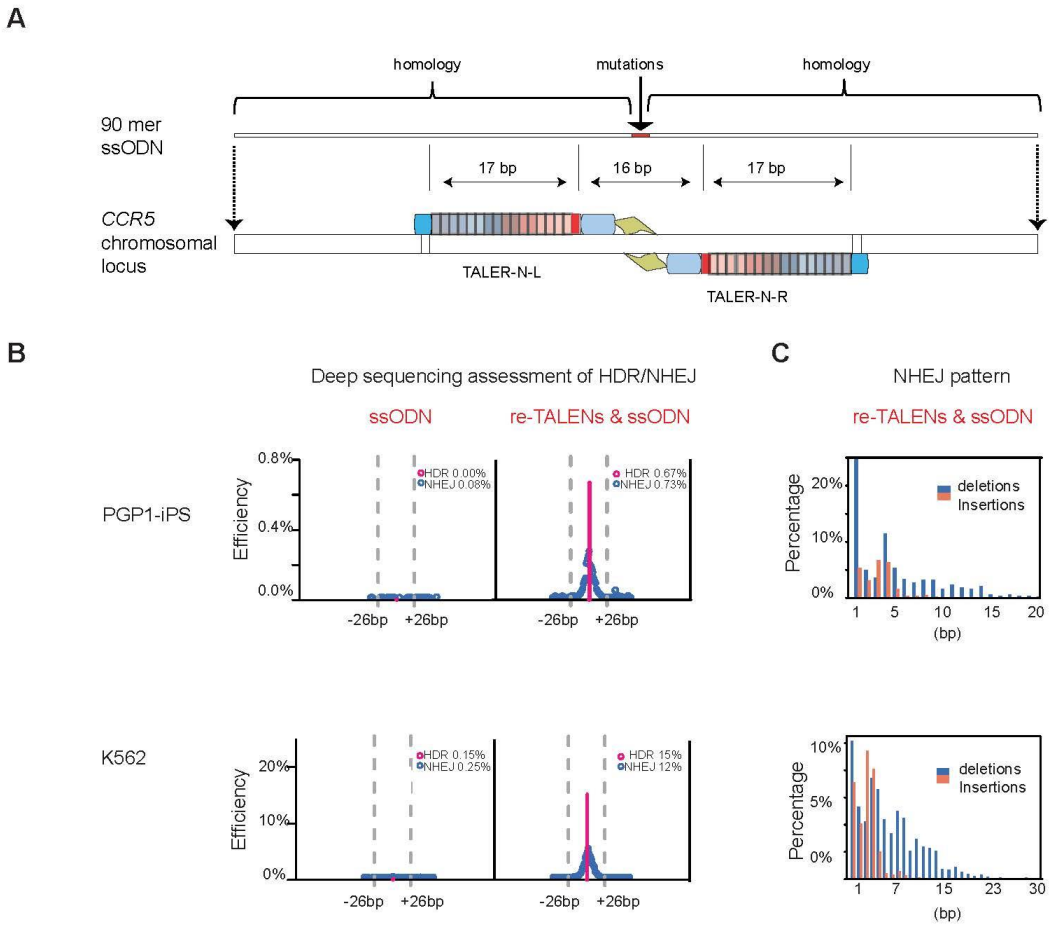


Figure 2_9. Sensitive and comprehensive genome editing assessment system (GEAS)
(A) Schematic representation of the genome engineering experimental design. At the re-TALEs pair targeting site, a 90mer ssODN carrying a 2bp mismatch against genomic DNA was delivered along with re-TALEN constructs into PGP1 iPSCs and K562 cells.
(B) Deep sequencing analysis of HDR and NHEJ efficiencies for re-TALEN pair and ssODN #9 from Table S3. Alterations in the genome of hiPSCs (top panel) and K562 cells (bottom panel) were analyzed from high-throughput sequence data by GEAS. HDR was quantified from the fraction of reads that contained a 2bp point mutation built into the center of the ssODN (pink), and NHEJ activity (blue) was quantified from the fraction of deletions/Insertions. We delivered the ssODN DNA donor alone (left panels) to the two cell types as the control. The gray dash lines mark the outer boundary of the re-TALEN pair's binding sites, which are at positions -26bp and +26bp relative to the center of the two re-TALEN binding sites. NHEJ-mediated genomic deletion frequencies at each nucleotide position are plotted in blue, HDR frequency is plotted in pink
. All reads analyzed as containing NHEJ insertions and deletions are included in the NHEJ percentage quoted in the figure, but only deletions are profiled in the graph. See Methods for details.
(C) Deletion/Insertion size distribution in hiPSCs (top) and K562 cells (Bottom) analyzed from the entire NHEJ population.

median deletion size of 6bp and insertion of 3bp in K562 cells (Figure 2_9 C), consistent with DNA lesion patterns usually generated by NHEJ (*40*). GEAS provides a convenient and accurate view of both NHEJ and HDR genome editing efficiencies. Moreover, unlike mismatch-endonuclease-based measures of NHEJ, our sequencing-based analysis gives a precise and immediate measure of the size profile of NHEJ indels generated by the nuclease-induced DSB. Finally, with GEAS, we observed levels of gene editing events in the hiPSCs that would be at or below the limit of detection of mismatch sensitive endonucleases (3%) (*41*).

We undertook several analyses to estimate the specificity, sensitivity and reproducibility of our platform. First, using the dataset of re-TALENs/ssODN experiment in hiPSCs, we computed the probabilities of observing the 2bp mismatch that signifies HDR by sequencing errors. Assuming that errors in different nucleotide positions and in forward versus reverse reads are independent, the probabilities for observed datasets were $<= 10^{-4}$ even for seeing a single HDR event among 7 $X10^5$ reads (Methods). These results indicate an extremely low false positive rate for HDR detection such that even seeing a single read with the targeted 2bp mismatch is a strong evidence of the presence of an HDR event. However this does not indicate the numerical accuracy of the measured HDR rate, or estimate the minimal HDR rate that could be reliably detected using this method. To estimate the latter, we performed an information-based analysis (Figure 2_10 A) that computationally spiked 'corrected' sequences into real reads, and assessed the signal of corrected reads relative to noise, namely the occurrence of any two bp mutations outside the target site (Methods). We found the HDR detection limit to be ~.007% for the hiPSCs data set, ~100 times lower than 0.67% HDR detected. Thereafter,

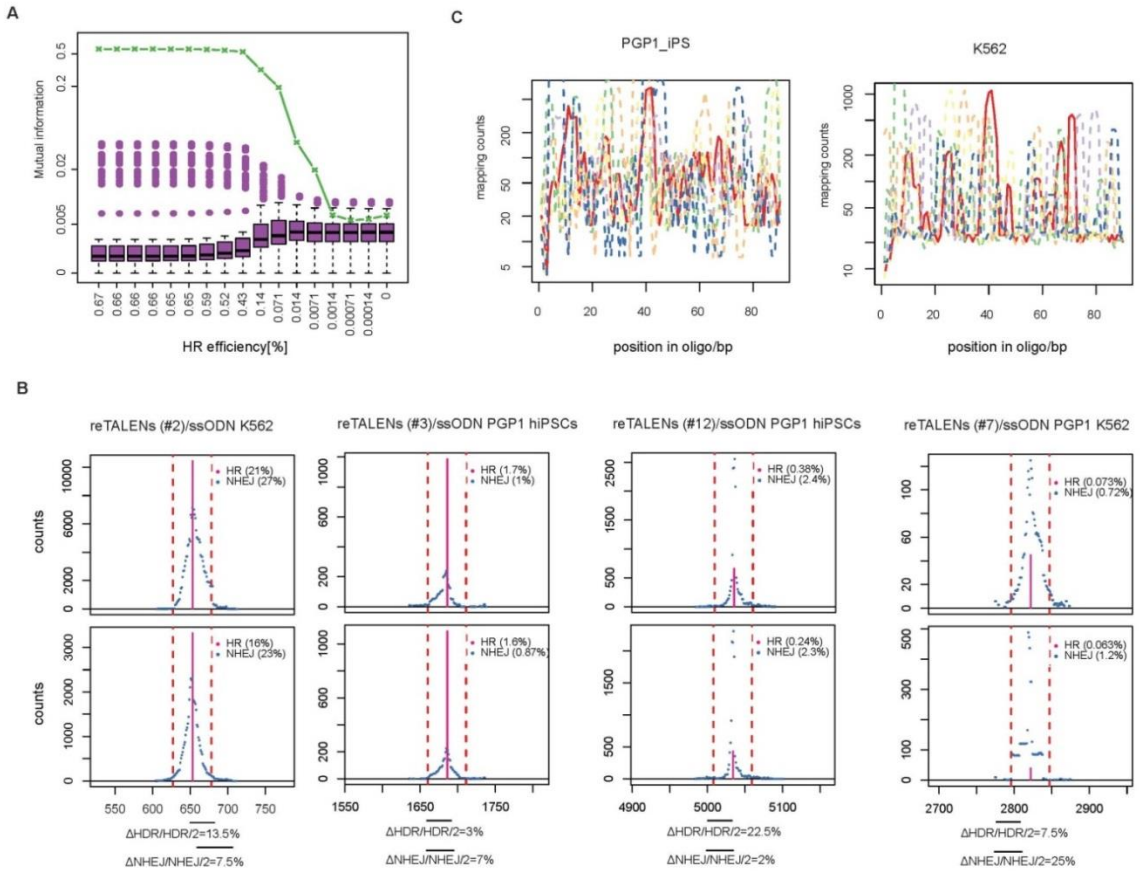


Figure 2_10. GEAS sensitivity and reproducibility test

(A) Information-based analysis of HDR detection limit. Given the dataset of re-TALENs (#3)/ssODN, we identified the reads containing the expected editing (HDR) and systematically removed these HDR reads to generate different artificial datasets with a "diluted" editing signal. We generated datasets with 100, 99.8, 99.9, 98.9, 97.8, 89.2, 78.4, 64.9, 21.6, 10.8, 2.2, 1.1, 0.2, 0.1, 0.02, and 0% removal of HDR reads to generate artificial datasets with HR efficiency ranging from 0~0.67%. For each individual dataset, we estimated mutual information (MI) of the background signal (in purple) and the signal obtained in the targeting site (in green). We observe that MI at the targeting site is remarkably higher than the background when the HDR efficiency is above 0.0014%. We estimated a limit of HDR detection between 0.0014% and 0.0071%. MI calculation is described in the Methods.

(B) The test of reproducibility of genome editing assessment system. The pairs of plots (Top and Bottom) show the HDR and NHEJ assessment results of two replicates with re-TALENs pair and cell type indicated above. For each experiment, we conducted nucleofection, targeted genome amplification, deep-sequencing and data analysis independently. We calculated the genome editing assessment variation of replicates as (|HDR1-HDR2|)/2 /((HDR+HDR2)/2) =ΔHDR/HDR and (|NHEJ1-NHEJ2|) /((NHEJ1+NHEJ2)/2) =ΔNHEJ/NHEJ and listed the variation results below the plots. We calculated the average variation of our system by (13.5%+7.5%+3%+7%+22.5%+2%+7.5%+25%)/8=10%. Factors that may contribute to the variations include the status of cells under nucleofection, nucleofection efficiency, and sequencing coverage and quality.

we performed the information-based measure of sensitivity for every individual data set to examine the reliability of our HDR measurement. Finally, over the course of our work, we performed replicate genome editing assessment analyses of four samples (Figure 2_10 B), finding that the relative degree of variation to be ~10% between the replicates. Among these datasets can be found two replicate measures of HDR at levels of ~0.07% (Figure 2_10 B), indicating that observation of rates of this magnitude are reproducible. The information-based measure of sensitivity and the *p*-value computation for specificity have been built into the R package for GEAS. An additional test of the accuracy of GEAS can be found below in our comparison of HDR rates measured from sequencing against rates measured from cloning (Figure 2_11 B, C).

Platform for isolation and clonal outgrowth of precisely edited hiPSCs

GEAS revealed that the re-TALEN pair #9 achieved precise genome editing with an efficiency of ~0.6~0.7% in hiPSCs, a level at which correctly edited cells can usually be isolated by growing out single cells in a few 96 well plates. Outgrowth of hiPSCs from single cells is generally difficult, but protocols were recently published that describe media that facilitate this procedure (*42*). We optimized these protocols along with single-cell sorting procedures to establish a robust platform for single hiPSCs sorting and maintenance, that enables scalable monoclonal hiPSC recovery with efficiencies of >25% (see Methods). We combined this with a rapid and efficient genotyping system with which we can conduct chromosomal DNA extraction outgrowths of our sorted, edited hiPSCs to be genotyped on a large scale. Together these components comprise a pipeline for robustly obtaining genome-edited hiPSCs without selection.

To demonstrate this pipeline (Figure 2_11 A), we first transfected PGP1 hiPSCs with a pair of re-TALENs and an ssODN targeting CCR5 at site #3 (see Table S3) and we performed GEAS with a portion of the transfected cells, finding an HDR frequency of 1.7% (Figure 2_11 B). This information along with the 25% recovery efficiency allowed us to estimate that we could obtain at least one single cell clone of correctly edited cells from five 96-well plate with Poisson probability 98% (assuming $\mu = 0.017 * 0.25 * 96 * 5 * 2$). We then FACS-sorted transfected single cells into 5 96-well plates 6 days after transfection and screened 100 monoclonal hiPSCs 8 days after sorting. Sanger sequencing revealed that 2 out of 100 of these unselected hiPSC colonies contained a heterozygous genotype possessing the 2bp mutation introduced by the ssODN donor (Figure 2_11 C). The efficiency (1%=2/2*100) was consistent with the next-generation sequencing analysis (1.7%) (Figure 2_11 B). The pluripotency of the resulting hiPSCs was confirmed with immunostaining for SSEA4 and TRA-1-60 (Figure 2_11 D). The cloned hiPSCs with desired genome editing generated mature teratomas with features of all three germ layers (Figure 2_11 E). To our knowledge, this is the first demonstration of using TALENs and ssODNs to obtain monoclonal hiPSCs with specific and scarless genetic alterations without any selection.

Application of the toolkit to alteration of 15 CCR5 *cis* sites

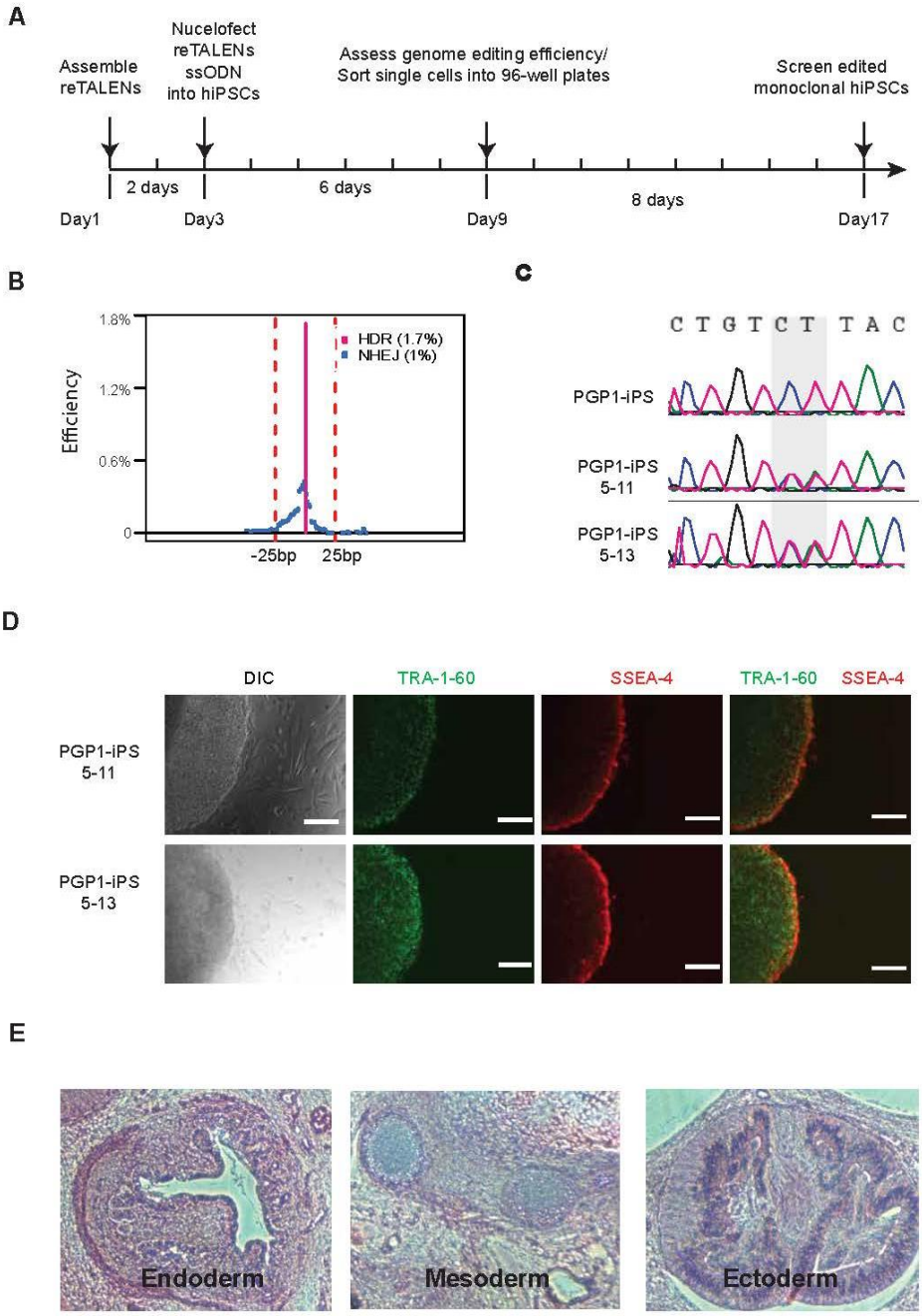


Figure 2_11. Using re-TALENs/ssODN to obtain monoclonal genome edited human iPS cells without selection
(A) Timeline of the experiment.
(B) Genome engineering efficiency of re-TALENs pair and ssODN (#3) assessed by the NGS platform described in Figure 5B
(C) Sanger sequencing results of monoclonal hiPS colonies after genome editing. Of note, the 2bp heterogeneous genotype (CT/CT→TA/CT) was successfully introduced into the genome of PGP1-iPS-5-11, PGP1-iPS-5-13 colonies.
(D) Immunofluorescence staining of targeted PGP1-iPS-5-11. Cells were stained for the pluripotency markers Tra-1-60 and SSEA4.
(E) Hematoxylin and eosin staining of teratoma sections generated from monoclonal PGP1-iPS-5-11 cells.

We next sought to test the scalability of our tools for building and assessing re-TALENs on 15 targeting sites *cis* to the CCR5 gene (Figure 2_12 A, Table S3). Anticipating that editing efficiency might depend on chromatin state, these sites were selected to represent a wide range of DNaseI sensitivities (*43*). Using our design and assembly tools above, we generated the re-TALEN pairs targeting these sites and transfected them with corresponding ssODNs into PGP1 hiPSCs and K562 cells. Six days after transfection, we profiled the genome editing efficiencies at these sites in both cell lines.

We detected NHEJ and HDR at levels above our statistical detection and sensitivity thresholds for 13/15 of our re-TALEN pair and ssODN in both hiPSC and K562 (see Table S4). In addition, in more than half of the re-TALENs pairs, the measured efficiency of HDR was > 0.3% in hiPSC and > 3% in K562 cells. Despite the fact that HDR & NHEJ efficiencies in K562 cells are on average 21X higher than those in hiPSCs (Figure 2_12 B,C), we observed many similarities across the cell types. A large and statistically significant positive 0.66 Pearson correlation coefficient is found between HR and NHEJ efficiency at the same targeting loci in both cell types (P=8X 10^{-5}) (Figure 2_12 D), consistent with the hypothesis that DSB generation, the common upstream step of both HDR and NHEJ, is a rate limiting step for genome editing. HDR rates are also very strongly correlated between the cell types (r=0.913, P= 2X10^{-6}). Factors contributing to the ~21x difference in rates between the cell types might include lower expression level of re-TALEs in hiPSCs (Figure 2_13), or the activity of DNA repair pathways. While a strong correlation in DNase I hypersensitivity (DNaseI HS) (*43*) between the cell types at the target sites (r=0.732, p=0) indicates similar chromatin states,

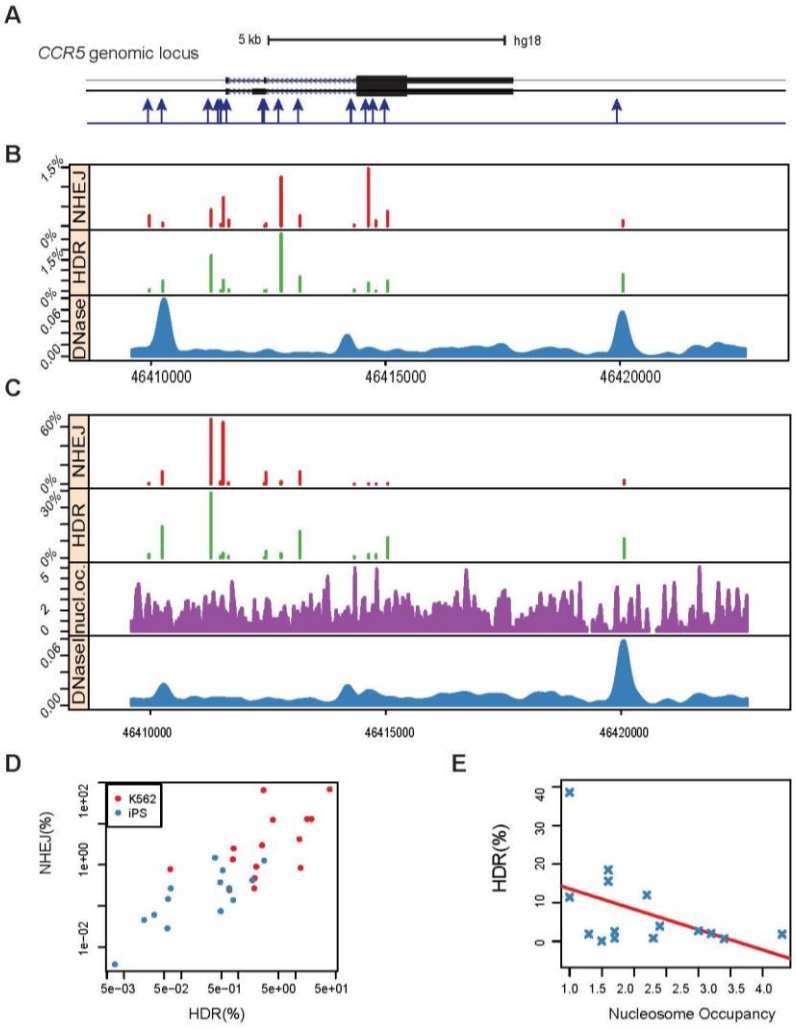


Figure 2_12. Re-TALENs/ssODNs genome editing activity on CCR5 in hiPSCs and K562 cells

(A) Schematic representation of the targeted genome editing sites on CCR5. The 15 targeting sites are illustrated by blue arrows below. For each site, cells were co-transfected with a pair of re-TALENs and its corresponding ssODNs donor carrying 2bp mismatch against the genomic DNA. The genome editing efficiencies were assayed 6 days after transfection.

(B) The HDR and NHEJ efficiencies of 15 pairs of re-TALENs/ssODNs targeting CCR5 in PGP1 hiPSCs genome. Top, NHEJ efficiencies were calculated by the frequency of genomic alleles carrying deletions or insertions at the targeting region. Middle, HDR efficiencies were calculated by the frequency of genomic alleles carrying 2bp mismatch introduced by ssODNs. Bottom, the DNaseI HS profile of hiPS cell line from ENCODE database (Duke DNase HS, iPS NIHi7 DS). The X-axis indicates the corresponding genomic position on chromosome 3.

(C) The HDR and NHEJ efficiencies of 15 pairs of re-TALENs/ssODNs on CCR5 in K562 cells. Panel1, NHEJ efficiencies were calculated by the frequency of genomic alleles carrying deletions or insertions at the targeting region. Panel2, HDR efficiencies were calculated by the frequency of genomic alleles carrying the 2bp mismatch introduced by ssODNs. Panel3, the nucleosome occupancy data of K562 cells from ENCODE (Stanf Nucleosome K562 Sig). Panel 4, DNaseI HS profile of K562 cells from ENCODE (Duke DNase HS K562 DS). The X-axis indicates the corresponding genomic position on Chromosome 3

(D) The correlation of HR and NHEJ efficiencies at identical sites in both iPSCs and K562 cells (r=0.68, P=8X 10-5).

(E) The inverse correlation of HR and nucleosome occupancy in K562 cells (r=-0.48, P=0.03)

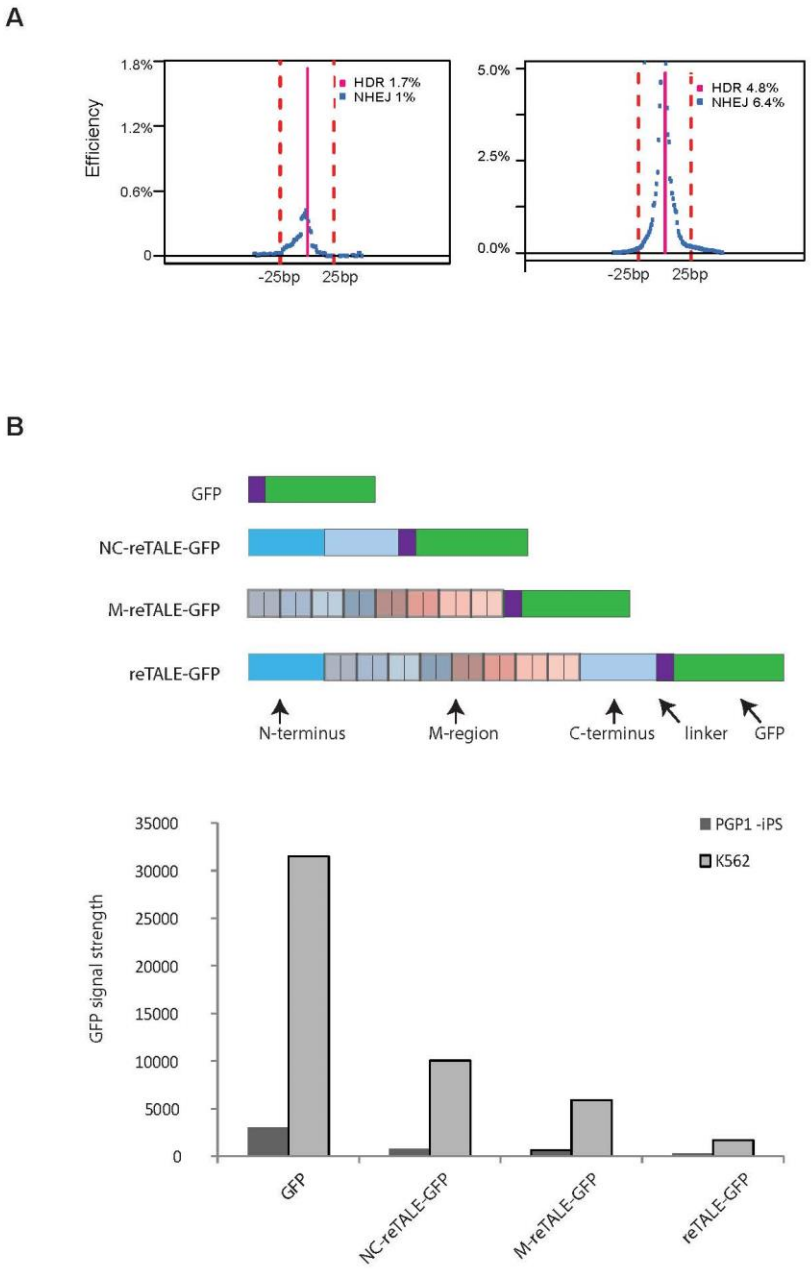


Figure 12_13. reTALE expression level and acitivities in K562 cells and hiPSCs

(A) Increased dosage of re-TALENs/ssODN increase the genome editing efficiency. We delivered re-TALENs pair (#3)/ssODNs into PGP1 hiPSCs with standardized condition (Top) (1 μg of each re-TALENs plasmid, 2 μl of 100 μM ssODN) or twice as much of re-TALENs and ssODN (Bottom) (2 μg of each re-TALENs plasmid, 4 μl of 100 μM ssODN). We observed increased HDR and NHEJ efficiency when we doubled the amount of DNA, although severe cell death was also observed in this group.

(B) The top panel illustrates the constructs we tested in the expression experiment. NC-re-TALE-GFP is a truncated form of re-TALE-GFP without the middle DNA recognition domain. M-re-TALE-GFP is a truncated form of re-TALE-GFP without the re-TALE's N and C termini. GFP is the control plasmid only encoding GFP. All the constructs were built in the same vector backbones. The bar graph indicates GFP expression level as measured by flow cytometry.

this correlation does not exclude the possibility of large differences in the scale of DNaseI HS between the cell types, which could thus also contribute to the 21x difference. The differences in HDR rates across sites in the same cell type may be attributed to different binding affinities between the re-TALENs and their target sites (*22*, *44*) or, again, to epigenetic status. Interestingly, while we did not observe any correlation between the genome editing efficiency and DNaseI HS (Figure 2_12 B,C Figure 2_14), we did observe an inverse correlation (r=-0.48, P=0.03) between nucleosome occupancy and HDR rates in K562 cells (Figure 2_12 E), for which (unlike hiPSC) these data are available (ENCODE/Stanford/BYU, (*45*)).

In this set of experiments, we leveraged the multiplicity and scalability of our methods to efficiently synthesize and assess genome editing at 15 sites *cis* to *CCR5* gene in both hiPSCs and K562s. We assembled re-TALENs in parallel in one 96-well plate using the one-hour TASA assembly reactions, and constructs were delivered into cell lines in parallel using 16-well nucleofector strips (Methods). Subsequently, we amplified targeting regions directly from cells in a 96-well plate using single-tube thermocycle reactions (Methods); and barcoded and pooled samples together for MiSeq sequencing runs.

Use of the toolkit to assess factors affecting genome editing with ssODNs

While ssODNs have been found to be effective as donor DNA in genome editing (see above, (*29*, *30*)), the mechanisms by which they participate in HDR are not understood and many questions remain regarding how to optimize their performance. To date ssODN

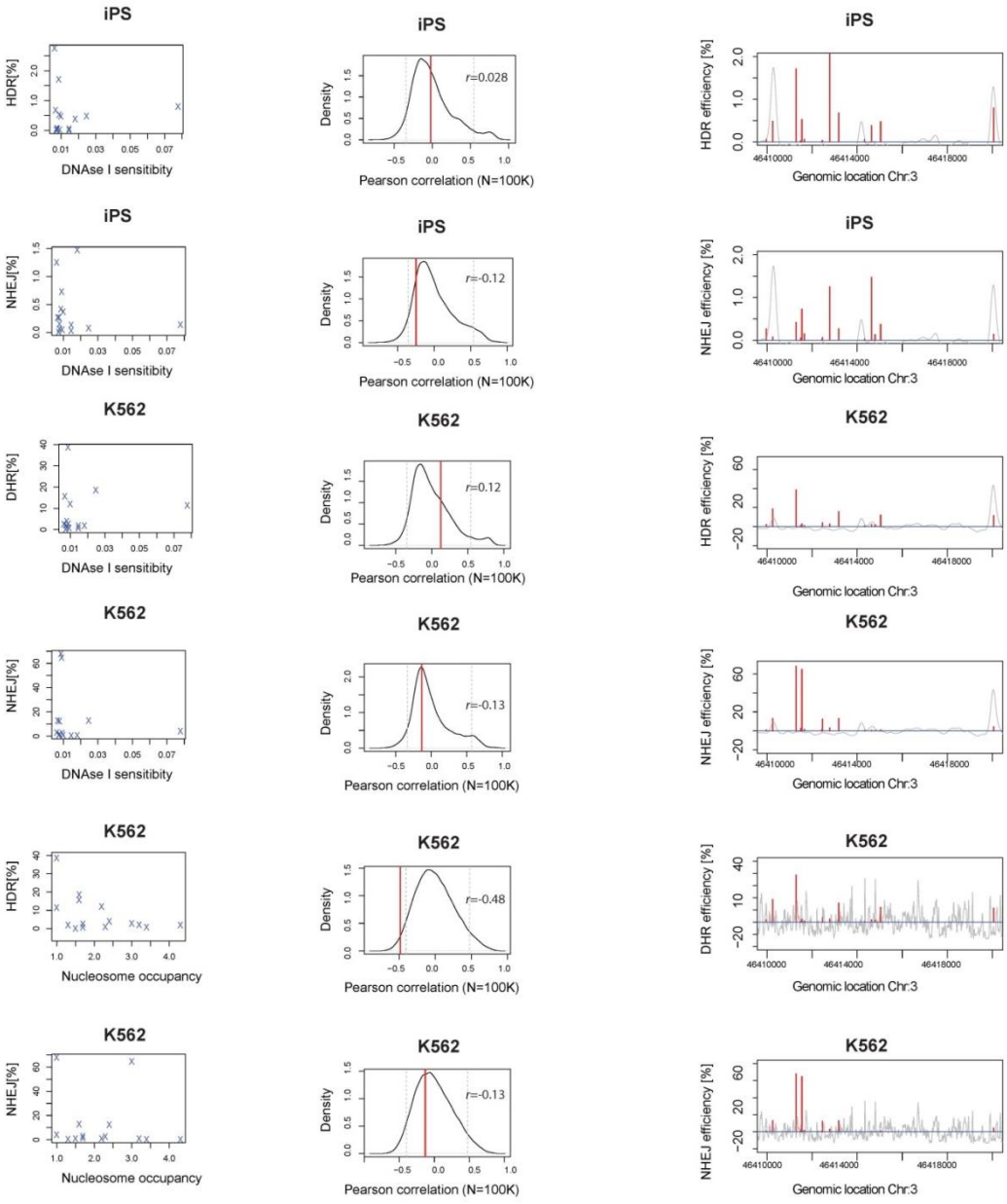


Figure 12_14. The correlation analysis of genome editing efficiency and epigenetic state
Left panel: we plotted the genome editing efficiencies (HDR or NHEJ) with the epigenetic parameters (DNaseI HS marker or Nucleosome occupancy).
Middle panel: we used Pearson correlation to study possible associations between epigenetic parameters (DNase I sensitivity, nucleosome occupancy) and genome engineering efficiencies (HR, NHEJ). We compared the observed correlation to a randomized set (N=100000). Observed correlations higher than the 95th percentile, or lower than the 5th percentile of the simulated distribution were considered as potential associations. We observed remarkable correlation between nucleosome occupancy and HR efficiency in K562 cells (r=-0.47).
Right panel: overlay of epigenetic parameters (DNase I sensitivity, nucleosome occupancy) and genome engineering efficiencies (HR, NHEJ) along the CCR5 in the genome.

incorporation has been assayed by methods that give limited information about incorporation outcomes, including mismatch sensitive endonuclease assays, generation of restriction sites (Chen et.al. 2011), and low throughput cloning and sequencing (*30*). To demonstrate the utility of the more comprehensive information provided by GEAS, we used it to analyze ssODN design and targeting parameters with re-TALENs in hiPSCs. First we designed a set of ssODNs of different lengths (50-170nt), all carrying the same 2bp mismatch in the middle of the spacer region of the CCR5 re-TALEN pair #3 target site. We observed that 90nt ssODN achieved the optimal HDR efficiency of ~1.8% and that longer ssODN declined in efficiency (Figure 2_15 A). Since it has been long established that longer homology regions improve HDR rates when dsDNA donors are used with nucleases (*46*), possible reasons for this result include that ssODNs are used in an alternative genome repair process, or longer ssODNs are less available to the DNA repair machinery, or that longer ssODNs incur negative effects that offset any improvements gained by longer homology, compared to dsDNA donors (*47*). However, if either of the former two hypothesis were the case, NHEJ rates would be unaffected with longer ssODNs. NHEJ induced insertion and deletion rates were observed to decline with HDR however (Figure 2_15 A), suggesting that the longer ssODNs present offsetting effects. Possible hypotheses would be that longer ssODNs are toxic to the cell (*48*), or that transfection of longer ssODNs saturates DNA processing machinery by causing decrease molar DNA uptake and so dilutes the capacity of the cells to take up or express re-TALEN plasmids.

Assuming that our 90nt ssODNs interact with genomic targeting site through homology directed pairing, we then explored whether there might be an optimal

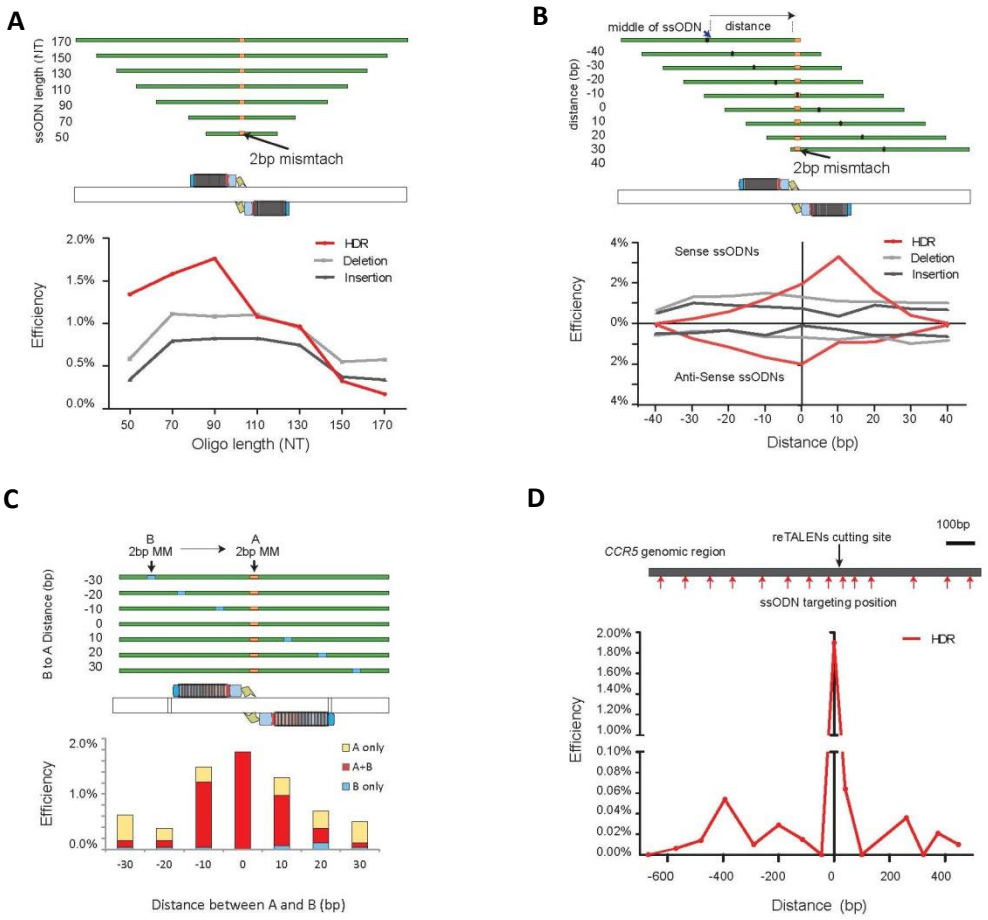


Figure 2_15. Functional parameters governing ssODN-mediated HDR with re-TALENs in iPSCs

(A) PGP1 iPSCs were co-transfected with re-TALENs pair (#3) and ssODNs of different lengths (50, 70, 90,110, 130,150,170 nts). All ssODNs possessed an identical 2bp mismatch against the genomic DNA in the middle of their sequence. A 90mer ssODN achieved optimal HDR in the targeted genome. The assessment of HDR, NHEJ-incurred deletion and insertion efficiency is described in the Methods.

(B) A set of 90nt sense and anti-sense ssODNs were delivered into PGP1 iPSCs with re-TALEN pair (#3). These ssODNs were designed so that the distance between the middle of ssODNs and the middle of re-TALENs cutting site varied from -40bp ~ 40bp while specifying the same 2bp mismatch against the genome at the center of the re-TALEN pair spacer region. ssODNs that were shifted to the 3' side of target site achieved higher HDR efficiencies than their 5' shifted counterparts; e.g., the sense ssODN at distance +10 (top half)achieved 3.3% HDR while that at distance -10 achieved 1.1% HDR. This asymmetry was also present when antisense ssODNs were used (bottom half).

(C) Ninety bp ssODNs corresponding to re-TALEN pair #3 each containing a 2bp mismatch (A) in the center and an additional 2bp mismatch (B) at different positions offset from A (where offsets varied from -30bp→30bp) were used to test the effects of deviations from homology along the ssODN. Genome editing efficiency of each ssODN was assessed in PGP1 hiPSC. The bottom bar graph shows the incorporation frequency of A only, B only, and A + B in the targeted genome. HDR rates decrease as the distance of homology deviations from the center increase

(D) ssODNs targeted to sites with varying distances (-620bp~ 480bp) away from the target site of re-TALEN pair #3 were tested to assess the maximum distance within which we can place ssODNs to introduce mutations. All ssODNs carried a 2bp mismatch in the middle of their sequences. We observed minimal HDR efficiency (<=0.06%) when the ssODN mismatch was positioned 40bp away from the middle of re-TALEN pair's binding site.

homology pairing arrangement of the ssODN design. To this end, we designed a set of 90nt sense ssODNs that maintained their identical 2bp mismatch against the center of the re-TALENs target site but asymmetrically shifted 5' or 3' with respect to the target region (Figure 2_15 B). We found that HDR rates remained elevated when the ssODNs were shifted to have more base pairing of their 3' ends against the complementary genomic DNA at the break compared to shifting an equivalent degree to have more pairing at the ssODN's 5' end. (Figure 2_15 B) However, this elevated HDR rate asymmetry vanishes when shifts reached ~25-30nt, at which point 5' and 3' shifts had similarly lower efficiencies. A similar asymmetrical preference for ssODN 3' pairing was seen when we used a set of antisense ssODNs. These results suggest that the HDR rates are improved when there is longer pairing up to 25-30nt of the 3' end of the ssODN against the chromosomal complement at the break, regardless of the strand polarity of the ssODN. This asymmetry of ssODN performance with placement accords with the structure of resected genomic DSBs (*40*), where 3' end ssDNA overhangs are exposed at DSB to ensure homologous recombination (HR). This suggests that ssODNs utilize resected genomic 3' overhangs DNA in a fashion similar to dsDNA donors to mediate genome editing. The fact that HDR rates drop off and become equivalent beyond that range may be due to the fact that increasing these shifts leaves less and less of the 5' end of the ssODN to pair against the indented chromosomal complement on the other side of the DSB, resulting in failure of the HDR. Notably, we see that NHEJ rates are unaffected by these shifts of the ssODN (Figure 2_15 B), suggesting that ssODN relative position and content do not have offsetting impacts on cellular NHEJ at a dsDNA break in the way that longer ssODNs appear to (above).

Next, we sought to examine the impact of imprecise homology in the pairing regions on HDR efficiency. Here we designed 90nt sense ssODNs all positioned symmetrically with

respect to the central 2bp mismatch (A) in the center of the spacer region of re-TALEN pair #3, where each also had a second 2bp mismatch (B) at a different offset from the center (Figure 2_15 C). A sense ssODN possessing only the center 2bp mismatch was used as a control. Each of these ssODNs was introduced individually with re-TALEN pair #3 and the outcomes were analyzed with GEAS. We found that overall HDR as measured by the rate at which the A mismatch was incorporated (A+B, A) decreased as the B mismatches increased their distance from the A mismatch (Figure 2_15 C, S7A). The higher HDR rate observed when B is only 10bp away from A may reflect a lesser need for pairing of the ssODN against genomic DNA proximal to the dsDNA break. We observe that at each distance of B to A, a fraction of HDR events only incorporate A, while another fraction incorporate both A and B (Figure 2_15 C (A and A+B)), These two events might also be interpretable in terms of gene conversion tracts (*49*) along the length of the ssDNA oligo, whereby A+B events represent long conversion tracts that extend beyond B, and A-only events represent shorter ones that do not reach to B. Under this interpretation, a distribution of gene conversion lengths in both directions along the ssODN can be estimated (Figure 2_16 B). The estimated distribution implies that gene conversion tracts progressively decrease in incidence as their lengths increase, a result very similar to gene conversion tract distributions seen with dsDNA donors, but on a highly compressed distance scale of tens of bp for the ssDNA donor vs. hundreds of bases for dsDNA donors. Consistent with this result, an experiment with an ssODN with three pairs of 2bp mismatches spaced at

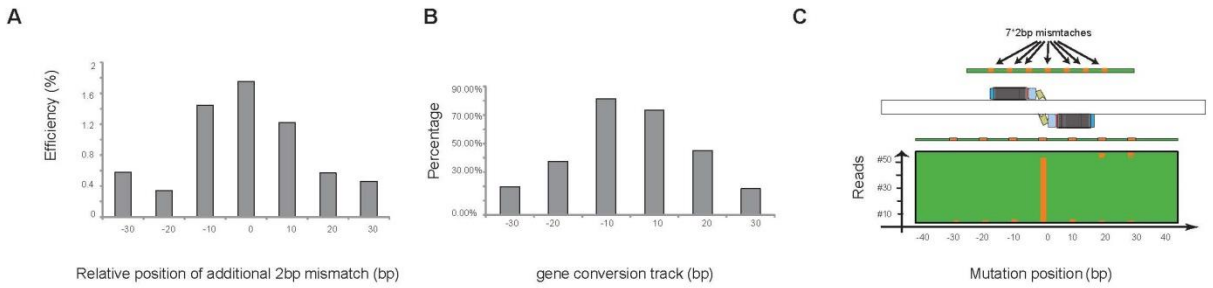


Figure 2_16. Study of ssODN designs and theoretical models underlying re-TALENs/ssODN mediated HDR
(A) Impact of homology pairing in the ssODN-mediated genome editing. Bar graph shows the rates of overall HDR as measured by the rate at which the middle 2b mismatch (A) was incorporated. Each bar represents the sum of the A+B and A values of the bars platted in Figure 2_15 C. Overall HDR decreases as the secondary mismatches B increase their distances from the A, where the relative position of B to A varies from -30bp to +30bp. The higher rates of incorporation when B is only 10bp away from A (-10bp and +10b) may reflect a lesser need for pairing of the ssODN against genomic DNA proximal to the dsDNA break.
(B) Distribution of gene conversion lengths along the ssODN. For each bar except the 0 bar in Figure 8C, the value (A+B)/(A + (A+B)) is interpreted as indicating the fraction of HDR events for which gene conversion extended at least as far as offset of B. These values are plotted here. The A-only events then represent shorter gene conversion tracts that do not extend as far as B. Gene conversion tracts progressively decrease in incidence as their lengths increase, a result very similar to gene conversion tract distributions seen with dsDNA donors (Elliott et al., 1998) but on a highly compressed distance scale of tens of bp for the ssDNA oligo vs. hundreds of bases for dsDNA donors.
(C) Assays for gene conversion tracts with dsDNA donors differ from experiments described in Figure 2_15 C and 2_16A, B by using a single dsDNA donor that contains a series of mutations and measuring contiguous series of incorporations (Elliott et al., 1998), whereas we used different ssODNs with single B mutations at different distances. Here, we used an ssODN donor with three pairs of 2bp mismatches spaced at intervals of 10nt on either side of the central 2bp mismatch A (Top). Genome editing with this ssODN gave rise of a pattern in which A alone was incorporated 85% (53/62) of the time, with multiple B mismatches incorporated at other times. Although numbers of B incorporation events were too low to estimate a distribution of tract lengths > 10bp, it is clear that the short tract region from -10-10bp predominates.

intervals of 10nt on either side of the central 2bp mismatch A gave rise of a pattern in which A alone was incorporated 86% of the time, with multiple B mismatches incorporated at other times (Figure 2_16 C). Although numbers of B incorporation events were too low to estimate a distribution of tract lengths > 10bp, it is clear that the short tract region from -10 to 10bp predominates (Figure 2_16 C). In none of these experiments do we see clear signs of an asymmetry indicating preference for incorporation of B mismatches on the 5' vs. the 3' side of the ssODN, as we saw above (Figure 2_15 B). Finally, in all of our experiments with single B mismatches, we see a small fraction of B-only incorporation events (0.04%~0.12%) that is roughly constant across all B distances from A (Figure 2_15 C). The nature of these events is unclear.

Finally, we sought to test how far we can place an ssODN away from the re-TALEN-induced dsDNA break and still observe incorporation, by delivering a set of 90nt ssODNs with central 2bp mismatches targeting a range of larger distances (-600bp-400bp) away from the re-TALEN-induced dsDNA break site. We observed >30x lower HDR efficiencies compared to the control ssODN positioned centrally over the cut region when the ssODNs matched ≥40bp away (Figure 2_15 D). This low level of incorporation distal to the DSB may be due to processes unrelated to the dsDNA cut, such as seen in experiments in which genomes are altered by ssDNA donor alone (*48*). Meanwhile, the low level of HDR seen at ~40bp may be due to the combination of weakened homology on the mismatch-containing side of the DSB coupled with insufficient ssODN oligo length on the other side of DSB that was also seen previously (Figure 2_15 B).

We observe that, in comparison with other methods of assessing design parameters for genome-editing, GEAS tool provided simultaneous information on rates of HDR, NHEJ, and

other mutagenic processes through a single experimental and statistical analysis method *vs.* performing different experiments and separate statistical analyses for each individually. Through the power of high-throughput sequencing, we could also detect events whose incidences were fractions of a percent, and successfully captured even lower frequency events by simply increasing the depth of sequencing. These features gave us the power to deduce that ssODN length might impact HDR rates through toxicity or saturation of DNA uptake capacity in the cell through a single set of similar experiments (Figure 2_15 A), and to detect possibly distinct processes of dsDNA cut-induced *vs.* non-cut induced oligo incorporation (Figure 2_15 D).

Discussion

Here we describe an efficient and integrated pipeline for the design, synthesis, and assessment of re-TALENs for human cell genome engineering in general, and for isolation of scarlessly engineered hiPSCs without selection in 17 days. The pipeline allowed us to address a number of challenging issues in genome-editing with TALENs. By eliminating DNA repeats, our recoded TALE design enabled one-hour, one-pot synthesis of TALENs, and allowed us to generate functional lenti-virus containing TALE sequences that will open the door to using re-TALEs and re-TALENs for a broad spectrum of cell types and *in vivo* models. The efficiency and scalability of our pipeline enabled us to investigate genome editing outcomes at 15 CCR5-proximal sites in parallel and explore correlations between editing rates and chromatin state in both hiPSCs and K562 cells. It also allowed us to explore multiple design parameters important for TALEN-based genome editing with ssODN donors with a single uniform experimental protocol and analysis method. The components in our pipeline are modular in that re-TALENs can be generated and used without applying GEAS, and the latter tools can be applied to targeted

genome-editing conducted by other means such as ZFNs, targeted nickases, meganucleases, and CRISPR systems (*50*). One can also simply use our recoded TALE sequences to design re-TALENs that can be ordered directly from gene synthesis companies instead. Our pipeline is also extensible. While our DNA repeat elimination algorithm was applied to a commonly used TALE RVD monomer and framework (*51*), it could just as well be applied to recoding the novel monomers (*44*, *52*) and frameworks (*33*). Finally, our pipeline is an open source. We provide complete details on the TASA protocol for one-pot re-TALE and re-TALEN assembly, are making our re-TALE sequences and plasmids available on Addgene, and are making software code and documentation for our recoding algorithm and GEAS system available to the scientific community. We envision that our pipeline will provide researchers with the means to facilitate and standardize their genome editing practice and extend them to additional cell lines and types.

With >100K reads / sample, GEAS was able to detect HDR events with efficiency as low as 0.007%, 400-fold more sensitive than the ~3% detection power provided by the mismatch-endonuclease method. Moreover, it provides genome alteration information at target sites at single nucleotide resolution, giving direct insight into NHEJ rates and indel profiles, as well as ssDNA oligo errors and other random mutations in a single step. We found that NHEJ lead to small (<10bp) deletions in both cell lines, consistent with the typical pattern of NHEJ products (*40*). We note, however, that since we only amplified ~250bp around the targeting sites in our analysis, we could not observe >200bp deletions reportedly found in U2OS cells (*24*). It has been reported that non-specific ssODN insertions are prevalent at DSBs in 293 cells (*31*). In our study, however, we found insertions generally occurred at very low incidence (Figure 2_10 C) and did not see evidence of overrepresentation of ssODN sequences (Figure 2_10 C), although the small size of the insertions we observed (median ~3bp) may have complicated this assessment. In this

study, we routinely pooled ~20 barcoded samples together and used the Illumina MiSeq system to obtain the sequence data which was analyzed with GEAS. Under optimized conditions, MiSeq can deliver >10 Million paired-end 150bp reads within 27hrs, so that up to 100 sample-barcoded targeting regions can be covered with ~100K reads each at a cost of approximately $10 per sample. If desired, sample throughput can be traded for higher sensitivity by reducing sample numbers and allotting more reads per sample. Higher capacity sequencing systems can be used to further improve throughput and cost per sample.

Our parallel analysis of genome editing in hiPSCs and K562 cells showed that despite the ~21x difference in editing rates between them, editing profiles were similar across 15 sites. Since gene targeting in hiPSCs is an important strategy for gene therapy, improving hiPSC editing rates is a high priority. The difference in re-TALE expression level (Figure 2_13 B), activity of DNA repair pathway and epigenetic status may contribute to the different genome editing efficiency between the two cell types. In the course of this study, we did, in fact, find that the increasing the amount of transfection constructs into the hiPSC increased HDR rates (Figure 2_13 A). In addition, we compared nucleosome occupancy data available for K562 cells (ENCODE/Stanford/BYU, (*45*)) with HDR rates in that cell type and we detected a strong correlation across targeting sites (Figure 2_12 E), suggesting the involvement of chromatin organization in the genome editing process. Such data will soon become available for hiPSC and give us a further opportunity to test these conclusions. Across both cell types, we also found that HDR and NHEJ rates were strongly correlated at all sites, which suggests that the rate of dsDNA cuts may be limiting for both processes. Contrary to the conventional conceptions that NHEJ is much more prevalent in human stem cells (*53*), we did not observe significant HDR/NHEJ ratio difference between K562 cells and hiPSCs (Figure 2_12 D).

We also used GEAS to explore design parameters of donor ssODN composition and targeting (Figure 2_15). While it has been long observed that gene targeting with long dsDNA donors is improved by long flanking homology regions (*46*), here we found that increasing ssODN length beyond 90 nt lowered HDR rates (Figure 2_15 A), possibly due to ssODN toxicity or saturation of DNA processing enzymes. However, we saw clear evidence of the need of ssODN to pair against both flanks of the dsDNA cut, with some preference for increased pairing of the 3' end of the ssODN. Results were symmetrical regardless of whether the ssODN was provided in a sense or an antisense orientation. Other properties of ssODN incorporation in HDR were similar to those of dsDNA donors, although operating on a compressed distance scale: For instance, we saw that deviations from homology at the flanks of the ssODN decreased incorporation (Figure 2_15 C), and also saw evidence that gene conversion tracts against ssODNs most frequently correspond to short regions around the dsDNA cut and extend outwards with smaller frequencies with increasing length (Figure 2_15 C, 2_16 B). We envision that further analyses of other design and protocol options for genome editing with ssODNs, such as use of phosphorothioated ssODNs and use of compounds that suppress viral response pathways (*48*), will lead to significant improvements in genome editing rates and outcomes and clarify the still obscure mechanism behind ssODN/nuclease mediated genome editing (Figure 2_17).

Finally, we note directions by which our pipeline may be extended. After clonal outgrowth and identification of correctly edited cells, the pipeline could be extended to interrogate phenotypes of these cells, and this could also be coupled to acquisition and analysis of transcriptomes of these cells obtained from RNA-seq analysis of barcoded aliquoted mixtures. In such experiments, cells found to have NHEJ alterations or no modifications could serve as useful comparisons and controls. Another direction would be to multiplex the introduction of

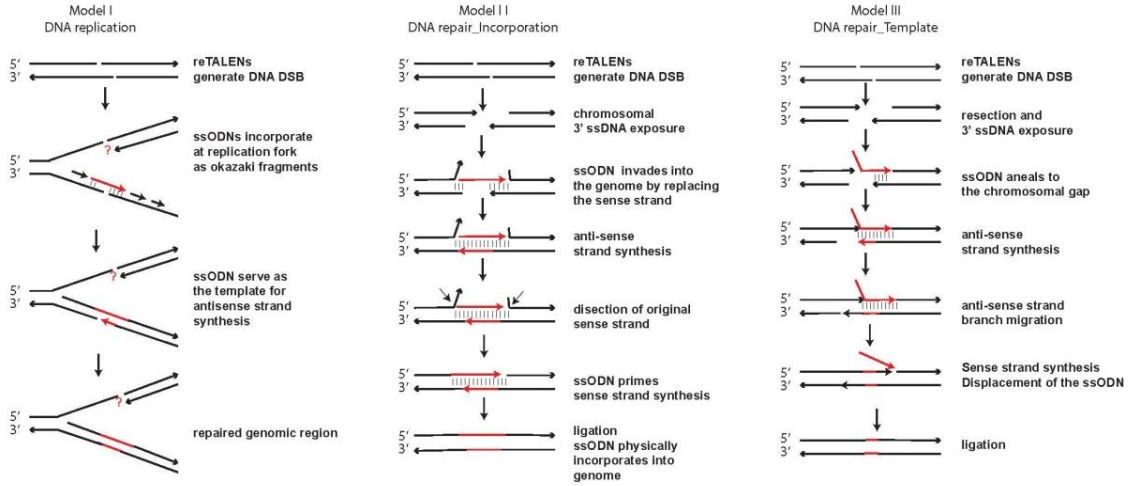


Figure 2_17. Theoretical models underlying re-TALENs/ssODN mediated HDR

Theoretical models underlying re-TALENs/ssODN mediated HDR. In model I, DNA DSB introduced by re-TALENs is not resolved until replication fork comes by, so that ssODN incorporates into the genome as Okazaki fragments to prime the synthesis of the nascent genomic DNA and also to serve as the template to repair the chromosomal complementary strand. However, in this model, the presence of the gap in the sense strand may cause the DNA replication fork (Question mark) to collapse.

In model II, DNA ends at DSB are digested by end processing enzymes so that 3’ ssDNA overhangs are generated at the re-TALENs cutting region. ssODN anneals to its chromosomal complement and invades into the genome by replacing the original strand. Subsequently, anti-sense strand is synthesized with the ssODN as the template and the replaced region at sense strand are resected. Finally, ssODN physically incorporates into the genome by the concerted action of DNA polymerases and ligases.

In model III, 3’ ssDNA overhangs are generated at the re-TALENs-mediated DSB as described in model II. The available chromosomal ssDNA region serves as the dock to recruit the binding of ssODN, which in turn serves as the template to initiate the repair of the anti-sense strand. Within the process of branch migration or template switching, the newly synthesized anti-sense strand re-anneals to the upstream genomic ssDNA overhangs and initiates the repair of sense strand. The newly synthesized sense strand displaces the ssODN from the genome and the genomic gap is sealed with the concerted action of DNA polymerases and ligases.

By examining our data sets, we first found both sense and anti-sense ssODNs, in cooperation with re-TALEs, are able to mediate genome editing (Figure 8B), which supports DNA repair models where ssODN strand polarity is not relevant. Consistent with this notion, ssODNs shifting to the 3’ end of the DSB achieves higher HDR efficiency than its 5’ counterparts (Figure 8B), mirroring the asymmetrical structure at DSB where 3’ chromosomal DNA overhangs are generated, potentially enabling the 3’ end of ssODNs to anneal. In addition, we observed a small window (~20bp) around DSBs where information encoded in the ssODNs can be passed to genomic DNA effectively, while genetic variations beyond this window cannot be effectively introduced into the genome (Figure 8C,D). The observation of this short conversion track (~20bp) with previous biotin-labeled oligo studies (Radecke et al., 2006) support model III where ssODNs function as template. It is conceivable that when ssODN serves as the template, it can only introduce genetic alteration into the chromosomal DNA at the DSB region, whereas the chromosomal overhangs preserve the flanking genomic information. However, it is also possible that ssODN physically incorporates into the genome but the mismatches beyond the original DSB region are repaired or resected. Additional research using radioactive-labeled ssODN and siRNA screening of important factors may further pinpoint these models.

reTALENs and/or ssODNs into the cells, and use the pipeline to generate libraries of single cells with different combinations of mutations within a set of sites, or of cells with a variety of different programmed mutations at one or more fixed sites. Such libraries would allow multilocus genetic influences on cellular phenotypes to be dissected or single nucleotide resolution of the bases in a regulatory element important for function. Longer term, a critical issue to applications of engineered hiPSCs, not addressed in the current pipeline, is determination of the presence of off-target mutations caused by use of nucleases to engineer cells. As costs of sequencing decline, we can envision conducting whole-genome sequencing of the monoclonal hiPSC that have been identified as correctly engineered at the target site to identify off-target edits. Using current technologies, ~50x coverage has been found to be sufficient to call single nucleotide variations in ~94% of the reference genome with a ~1% false positive rate (*54*), although detection of indels typically generated by NHEJ may less reliable (*55*). Shorter term, but less definitive, we could perform large-scale targeted sequencing of genomic sites that are similar in sequence to target sites or which have been identified as sites of off-target activity by *in vivo* (*56*) or *in vitro* (*57*) assays. Application of off-target detection to monoclonal outgrowths engineered cells has advantages over detection of off-targets in cell populations in that the latter is invariably limited by sequencing depth to detecting only relatively common off-targets, and does not reveal the distribution of high frequency *vs.* rare off-targets that might be found in single cells. Integration of these many directions for development of our platform for efficiently isolating scarlessly engineered human stem cells will give the research community many new abilities to analyze the causal underpinnings of numerous important biological problems, as well as methods to prepare hiPSC and other cell lines to precise specifications that could be useful for disease treatment.

Materials and Methods

re-TALEs design

re-TALEs were optimized at different levels to facilitate assembly, and improve expression. re-TALE DNA sequences were first co-optimized for a human codon-usage, and low mRNA folding energy at the 5' end (GeneGA, Bioconductor). The obtained sequence was evolved through several cycles to eliminate repeats (direct or inverted) longer than 11 bp. Specifically, the re-TALE sequence was evolved in several design cycles to eliminate repeats. In each cycle, synonymous sequences from each repeat are evaluated. Those with the largest hamming distance to the evolving DNA are selected. The final sequence with cai = 0.59 ΔG= -9.8 kcal/mol. We provide an R package to carry out this general framework for synthetic protein design. The sequence of one of re-TALE possessing 16.5 monomers is listed in Sequence 1. We provide an R package in the supplement to carry out this general framework for synthetic protein design.

re-TALE assembly

(1) re-TALE dimer blocks preparation

re-TALE dimer blocks encoding two RVDs were generated by two rounds of PCR under standard Kapa HIFI (KPAP) PCR conditions, in which the first round of PCR introduced the RVD coding sequence and the second round of PCR generated the entire dimer blocks with 36bp overlaps with the adjacent blocks. PCR products were purified using QIAquick 96 PCR

Purification Kit (QIAGEN) and the concentrations were measured by Nano-drop. The primer and template sequences are listed in Table 2_1 and 2_ 2.

(2) re-TALE destination vectors preparation

re-TALENs and re-TALE-TF destination vectors were constructed by modifying the TALE-TF and TALEN cloning backbones (*51*). We re-coded the 0.5 RVD regions on the vectors

Table 2_1. re-TALE blocks sequences

block0	CGCAATGCGCTCACGGGAGCACCCCTCAACCTAACCCCTGAACAGGTAGTCGCTATAGCTTCAN NNNNNGGGGGCAAGCAAGCACTTGAGACCGTTCAACGACTCCTGCCAGTGCTCTGCCAAGCCCA TGGATTGACTCCGGAGCAAGTCGTCGCGATCGCGAGCNNNNNNGGGGGGAAGCAGGCGCTGGAA ACTGTTCAGAGACTGCTGCCTGTACTTTGTCAGGCGCATGGTCTC
block1	AGACTGCTGCCTGTACTTTGTCAGGCGCATGGTCTCACCCCCGAACAGGTTGTCGCAATAGCAA GTNNNNNNGGCGGTAAGCAAGCCCTAGAGACTGTGCAACGCCTGCTCCCCGTGCTGTGTCAGGC TCACGGTCTGACACCTGAACAAGTTGTCGCGATAGCCAGTNNNNNNGGGGGAAAACAAGCTCTA GAAACGGTTCAAAGGTTGTTGCCCGTTCTGTGCCAAGCACATGGGTTA
block1'	TGCGCTCACGGGAGCACCCCTCAACCTCACCCCCGAACAGGTTGTCGCAATAGCAAGTNNNNNN GGCGGTAAGCAAGCCCTAGAGACTGTGCAACGCCTGCTCCCCGTGCTGTGTCAGGCTCACGGTC TGACACCTGAACAAGTTGTCGCGATAGCCAGTNNNNNNGGGGGAAAACAAGCTCTAGAAACGGT TCAAAGGTTGTTGCCCGTTCTGTGCCAAGCACATGGGTTA
block2	AGGTTGTTGCCCGTTCTGTGCCAAGCACATGGGTTAACACCCGAACAAGTAGTAGCGATAGCGT CANNNNNNGGGGGTAAACAGGCTTTGGAGACGGTACAGCGGTTATTGCCGGTCCTCTGCCAGGC CCACGGACTTACGCCAGAACAGGTGGTTGCAATTGCCTCCNNNNNNGGCGGGAAACAAGCGTTG GAAACTGTGCAGAGACTCCTTCCTGTTTTGTGTCAAGCCCACGGCTTGACGCCT
block3	AGACTCCTTCCTGTTTTGTGTCAAGCCCACGGCTTGACGCCTGAGCAGGTTGTGGCCATCGCTA GCNNNNNNGGAGGGAAGCAGGCTCTTGAAACCGTACAGCGACTTCTCCCAGTTTTGTGCCAAGC TCACGGGCTAACCCCCGAGCAAGTAGTTGCCATAGCAAGCNNNNNNGGAGGAAAACAGGCATTA GAAACAGTTCAGCGCTTGCTCCCGGTACTCTGTCAGGCACACGGTCTA
block4	CGCTTGCTCCCGGTACTCTGTCAGGCACACGGTCTAACTCCGGAACAGGTCGTAGCCATTGCTT CCNNNNNNGGCGGCAAACAGGCGCTAGAGACCGTCCAGAGGCTCTTGCCTGTGTTATGCCAGGC ACATGGCCTCACCCCGGAGCAGGTCGTTGCCATCGCCAGTNNNNNNGGCGGAAAGCAAGCTCTC GAAACAGTACAACGGCTGTTGCCAGTCCTATGTCAAGCTCATGGACTG
block5	CGGCTGTTGCCAGTCCTATGTCAAGCTCATGGACTGACGCCCGAGCAGGTAGTGGCAATCGCAT CTNNNNNNGGAGGTAAACAAGCACTCGAGACTGTCCAAAGATTGTTACCCGTACTATGCCAAGC GCATGGTTTAACCCCAGAGCAAGTTGTGGCTATTGCATCTNNNNNNGGTGGCAAACAAGCCTTG GAGACCGTGCAACGATTACTGCCTGTCTTATGTCAGGCCCATGGCCTT
block6	CGATTACTGCCTGTCTTATGTCAGGCCCATGGCCTTACTCCTGAGCAGGTGGTCGCTATCGCCA GCNNNNNNGGGGGCAAGCAAGCACTGGAAACAGTCCAGCGTTTGCTTCCAGTACTTTGTCAGGC GCATGGATTGACACCGGAACAAGTGGTGGCTATAGCCTCANNNNNNGGAGGAAAGCAGGCGCTG GAAACCGTCCAACGTCTTTTACCGGTGCTTTGCCAGGCGCACGGGCTC
block6'	CGATTACTGCCTGTCTTATGTCAGGCCCATGGCCTTACTCCTGAGCAAGTCGTAGCTATCGCCA GCNNNNNNGGTGGGAAACAGGCCCTGGAAACCGTACAACGTCTCCTCCCAGTACTTTGTCAAGC ACACGGGTTGACACCGGAACAAGTGGTGGCGATTGCGTCCNNNNNNGGAGGCAAGCAGGCACTG GAGACCGTCCAACGGCTTCTTCCGGTTCTTTGCCAGGCTCATGGGCTC
block7	CGGCTTCTTCCGGTTCTTTGCCAGGCTCATGGGCTCACGCCAGAGCAGGTGGTAGCAATAGCGT CGNNNNNNGGTGGTAAGCAAGCGCTTGAAACGGTCCAGCGTCTTCTGCCGGTGTTGTGCCAGGC GCACGGACTCACACCAGAACAAGTGGTTGCTATTGCTAGTNNNNNNGGTGGAAAGCAGGCCCTC GAGACGGTGCAGAGGTTACTTCCCGTCCTCTGTCAAGCGCACGGCCTC

Table 2_2. re-TALE blocks primer sequences

block0-F	CGCAATGCGCTCACGGGAGCACCCCTCAACctAACCCCTGAACAGGT*A*G
block0-R	GAGACCATGCGCCTGACAAAGTACAGGCAGCAGTCTCTGAACAG*T*T
block1'-F	TGGCGCAATGCGCTCACGGGAGCACCCCTCA*A*C
block1-F	AGACTGCTGCCTGTACTTTGTCAGGCGCATGGTCTCACCCCCGAACA*G*G
block1-R/block1'-R	TAACCCATGTGCTTGGCACAGAACGGGCAACAACCTTTGAACCG*T*T
block2-F	AGGTTGTTGCCCGTTCTGTGCCAAGCACATGGGTTAACACCCgaac*a*a
blcok2-R	AGGCGTCAAGCCGTGGGCTTGACACAAAACAGGAAGGAGTCTCTGCACAG*T*t
block3-F	AGACTCCTTCCTGTTTTGTGTCAAGCCCACGGCTTGACGCCTG*A*G
block3-R	TAGACCGTGTGCCTGACAGAGTACCGGGAGCAAGCGCT*G*A
block4-F	CGCTTGCTCCCGGTACTCTGTCAGGCACACGGTCTAA*C*T
block4-R	CAGTCCATGAGCTTGACATAGGACTGGCAACAGCCGTT*G*T
block5-F	CGGCTGTTGCCAGTCCTATGTCAAGCTCATGGACTGA*C*G
block5-R	AAGGCCATGGGCCTGACATAAGACAGGCAGTAATCGTT*G*C
block6-F	CGATTACTGCCTGTCTTATGTCAGGCCCATGGCCTTA*C*T
block6-R	GAGCCCGTGCGCCTGGCAAAGCACCGGTAAAAGACGTTGGA*C*G
block6'-F	CGATTACTGCCTGTCTTATGTCAGGCCCATGGCCTTACTCCTGAGCAA*G*T
block6'-R	GAGCCCATGAGCCTGGCAAAGAACCGGAAGAAGCCGTT*G*G
block7-F	CGGCTTCTTCCGGTTCTTTGCCAGGCTCATGGGCTCACGCCAGAGCAGG*T*G
blcok7-R	GAGGCCGTGCGCTTGACAGAGGACGGGAAGTAACCTCT*G*C

and also incorporated SapI cutting site at the designated re-TALE cloning site.. Plasmids can be pre-treated with SapI (New England Biolabs) with manufacturer recommended conditions and purified with QIAquick PCR purification kit (QIAGEN).

(3) TASA assembly

We carried out the (10ul) one-pot TASA assembly reaction with 200ng of each block, 500ng destination backbone, 1X TASA enzyme mixture (2U SapI, 100U Ampligase (Epicentre), 10mU T5 exonuclease (Epicentre), 2.5U Phusion DNA polymerase (New England Biolabs)) and 1X isothermal assembly reaction buffer as described before (*58*) (5% PEG-8000, 100 mM Tris-HCl pH 7.5, 10 mM MgCl2, 10 mM DTT, 0.2 mM each of the four dNTPs and 1 mM NAD). Incubations were performed at 37°C for 5min and 50 °C for 30 min. Alternatively, >90% efficiency can be achieved by two-steps assembly. First, 10ul re-TALE assembly reactions were performed with 200ng of each block, 1X re-TALE enzyme mixture (100U Ampligase, 12.5mU T5 exonuclease, 2.5U Phusion DNA polymerase) and 1X isothermal assembly buffer at 50°C for 30min, followed by standardized Kapa HIFI PCR reaction, agarose gel electrophoresis, and QIAquick Gel extraction (Qiagen) to enrich the full length re-TALEs. 200ng re-TALE amplicons can then be mixed with 500ng Sap1-pre-treated destination backbone, 1X re-TALE assembly mixture and 1X isothermal assembly reaction buffer and incubated at 50 °C for 30 min. TASA assembly reaction and re-TALE final assembly reaction can be processed directly for bacterial transformation to colonize individual assemblies.

<u>Assessment of re-TALEs functionality</u>

(1) Cell culture

PGP1 iPS cells were maintained on Matrigel (BD Biosciences)-coated plates in mTeSR1 (Stemcell Technologies). Cultures were passaged every 5–7 days with TrypLE Express (Invitrogen). 293T and 293FT cells were grown and maintained in Dulbecco's modified Eagle's medium (DMEM, Invitrogen) high glucose supplemented with 10% fetal bovine serum (FBS, Invitrogen), penicillin/streptomycin (pen/strep, Invitrogen), and non-essential amino acids (NEAA, Invitrogen). K562 cells were grown and maintained in RPMI (Invitrogen) supplemented with 10% fetal bovine serum (FBS, Invitrogen 15%) and penicillin/streptomycin (pen/strep, Invitrogen). All cells were maintained at 37°C and 5% CO2 in a humidified incubator.

(2) re-TALE-TF activity assessment

re-TALE-TF activity assessment experiments were conducted as described before (*26*). Briefly, 293T cells were seeded onto 24-well plates the day before transfection at densities of 2 × 105 cells well. Approximately 24 h after initial seeding, cells were co-transfected with 500ng plasmids carrying re-TALE-TF-2A-GFP and 30ng mCherry reporters using Lipofectamine 2000 (Invitrogen) following the manufacturer's protocols. Cells were harvested using TrypLE Express (Invitrogen) ~18 h after transfection and resuspended in 200 µl of media for flow cytometry analysis using an LSRFortessa cell analyzer (BD Biosciences). The flow cytometry data were analyzed using BD FACSDiva (BD Biosciences). At least 25,000 events were analyzed for each transfection sample.

(3) re-TALENs activity assessment

We established a stable 293T cell line for detecting HDR efficiency as described before (*20*). Specifically, the reporter cell lines bear genomically integrated GFP coding sequences disrupted by the insertion of a stop codon and a 68bp genomic fragment derived from the AAVS1 locus. We seeded reporter cells at densities of 2 × 105 cells per well in 24-well plate and transfected them with 1μg of each re-TALENs plasmid and 2μg DNA donor plasmid using Lipofectamine 2000 following the manufacturer's protocols. Cells were harvested using TrypLE Express (Invitrogen) ~18 h after transfection and resuspended in 200 μl of media for flow cytometry analysis using an LSRFortessa cell analyzer (BD Biosciences). The flow cytometry data were analyzed using FlowJo (FlowJo). At least 25,000 events were analyzed for each transfection sample. For endogenous AAVS1 locus targeting experiment in 293T, the transfection procedures were identical as described above and we conducted puromycin selection with drug concentration at 3μg/ml 1 week after transfection.

(4) Functional lentivirus generation assessment

The lentiviral vectors were created by standard PCR and cloning techniques. The lentiviral plasmids were transfected by Lipofectamine 2000 with Lentiviral Packaging Mix (Invitrogen) into cultured 293FT cells (Invitrogen) to produce lentivirus. Supernatant was collected 48 and 72h post-transfection, sterile filtered, concentrated with the Amicon Ultra-15 Centrifugal Filter Units (Millipore) and added at different dilutions to fresh 293T cells with polybrene. Cells were harvested using TrypLE Express (Invitrogen) and resuspended in 200 μl of media for flow cytometry analysis using an LSRFortessa cell analyzer (BD Biosciences). The flow cytometry data were analyzed using BD FACSDiva (BD Biosciences). At least 25,000 events were analyzed for each transfection sample. Lentivirus titration was calculated based on

the following formula: virus titration = (percentage of GFP+ 293T cell * initial cell numbers under transduction) / (the volume of original virus collecting supernatant used in the transduction experiment). To test the functionality of lentivirus, 3 days after transduction, we transfected 2 × 105 lentivirus transduced 293T cells with 30 ng plasmids carrying mCherry reporter and 500ng pUC19 plasmids using Lipofectamine 2000 (Invitrogen). Cell images were analyzed using Axio Observer Z.1 (Zeiss) 18 hours after transfection and harvested using TrypLE Express (Invitrogen) and resuspended in 200 µl of media for flow cytometry analysis using a LSRFortessa cell analyzer (BD Biosciences). The flow cytometry data were analyzed using BD FACSDiva (BD Biosciences).

Test of re-TALENs/ssODNs genome editing efficiency

(1) re-TALENs/ssODNs nucleofection in PGP1 hiPSCs and K562 cells

PGP1 iPSCs were cultured in Rho kinase (ROCK) inhibitor Y-27632 (Calbiochem) 2h before nucleofection. Transfections were done using P3 Primary Cell 4D-Nucleofector X Kit (Lonza). Specifically, cells were harvested using TrypLE Express (Invitrogen) and 2×106 cells were resuspended in 20 µl nucleofection mixture containing 16.4 µl P3 Nucleofector solution, 3.6 µl supplement, 1µg of each re-TALENs plasmid, 2µl of 100 µM ssODN. Subsequently, we transferred the mixtures to 20µl Nucleocuvette strips and conducted nucleofection using CB150 program. Cells were plated on Matrigel-coated plates in mTeSR1 medium supplemented with ROCK inhibitor for the first 24 hrs. For endogenous AAVS1 locus targeting experiment with dsDNA donor, we utilized the identical procedure except we used 2 µg dsDNA donor and we supplement the mTeSR1 media with puromycin at the concentration of 0.5ug/mL 1 week after transfection.

K562 cells were nucleofected with re-TALENs/ssODNs using SF Cell Line 4D-Nucleofector X Kit. Specifically, 2×106 cells were resuspended in 20 μl nucleofection mixture containing 16.4 μl SF Nucleofector solution, 3.6 μl supplement, 1 μg of each re-TALENs plasmid, and 2 μg of corresponding ssODN donor. Subsequently, we transferred the mixtures to 20 μl Nucleocuvette strips and conducted nucleofection using FF120 program. Cells were transferred to pre-warmed medium.

The information of reTALENs and ssODNs used in this study are listed in Table 2_3 to 2_6.

(2) Amplicon library preparation of the targeting regions

Cells were harvested 6 days after nucleofection and 0.1 μl prepGEM tissue protease enzyme (ZyGEM) and 1 μl prepGEM gold buffer (ZyGEM) were added to 8.9 μl of the 2~5 X 105 cells in the medium. 1ul of the reactions were then added to 9 μl of PCR mix containing 5ul 2X KAPA Hifi Hotstart Readymix (KAPA Biosystems) and 100nM corresponding amplification primer pairs. Reactions were incubated at 95°C for 5 min followed by 15 cycles of 98°C, 20 s; 65°C, 20 s and 72°C, 20 s. To add the Illumina

Table 2_3. Information of re-TALEN pairs/ssODN targeting CCR5

# targeting site	re-TALENs pair targeting site (start) /chr3:	re-TALENs pair targeting site (end) /chr3:	re-TALEN-L targeting sequence	re-TALE-R targeting sequence	ssODN donor sequence
1	46409942	46409993	TCCCCACTTTCTTGTGAA	TAACCACTCAGGACAGGG	CTGAAGAATTTCCCATGGGTCCCCACTTTCTTGTGAATCCTTGGAGTGAACCCCCCTGTCCTGAGTGGTTACTAGAACACACCTCTGGAC
2	46410227	46410278	TCACACAGCAAGTCAGCA	TAGCGGAGCAGGCTCGGA	TGGAAGTATCTTGCCGAGGTCACACAGCAAGTCAGCAGCACAGCCAGTGTGACTCCGAGCCTGCTCCGCTAGCCCACATTGCCCTCTGGG
3	46411260	46411311	TACCCAGACGAGAAAGCT	TCAGACTGCCAAGCTTGA	CTACTGTCATTCAGCCCAATACCCAGACGAGAAAGCTGAGGGTATAACAGGTTTCAAGCTTGGCAGTCTGACTACAGAGGCCACTGGCTT
4	46411464	46411515	TCTTGTGGCTCGGGAGTA	TATTGTCAGCAGAGCTGA	GGAAGCCCAGAGGGCATCTTGTGGCTCGGGAGTAGCTCTCTGCTACCTTCTCAGCTCTGCTGACAATACTTGAGATTTTCAGATGTCACC
5	46411517	46411568	TTGAGATTTTCAGATGTC	TATACAGTCATATCAAGC	TCAGCTCTGCTGACAATACTTGAGATTTTCAGATGTCACCAACGCCCAAGAGAGCTTGATATGACTGTATATAGTATAGTCATAAAGAAC
6	46411634	46411685	TTCAGATAGATTATATCT	TGCCAGATACATAGGTGG	GTGGAAAATTTCTCATAGCTTCAGATAGATTATATCTGGAGTGAGCAATCCTGCCACCTATGTATCTGGCATAGTGTGAGTCCTCATAAA
7	46412396	46412447	TTATACTGTCTATATGAT	TCAGCTCTTCTGGCCAGA	GAAACAGCATTTCCTACTTTTATACTGTCTATATGATTGATTTGGTCAGCTCATCTGGCCAGAAGAGCTGAGACATCCGTTCCCCTACAA
8	46412432	46412483	TGGCCAGAAGAGCTGAGA	TTACCGGGGAGAGTTTCT	TTGATTTGCACAGCTCATCTGGCCAGAAGAGCTGAGACATCCGTATCCCTACAAGAAACTCTCCCCGGTAAGTAACCTCTCAGCTGCTTG
9	46412750	46412801	TTTGCAGAGAGATGAGTC	TTAGCAGAAGATAAGATT	GGAGAGGGTTTAGTTCTCCTTAGCAGAAGATAAGATTTCAAGATGAGAGCTAAGACTCATCTCTCTGCAAATCTTTCTTTTGAGAGGTAA
10	46413152	46413203	TATAAGACTAAACTACCC	TCGTCTGCCACCACAGAT	TAATATAATAAAAAATGTTTCGTCTGCCACCACAGATGAATGTCGAGCATTCTGGGTAGTTTAGTCTTATAACCAGCTGTCTTGCCTAGT
11	46414305	46414356	TAAAACAGTTTGCATTCA	TATAAAGTCCTAGAATGT	TTAAAAACCTATTGATGTATAAAACAGTTTGCATTCATGGAGGGTGACTAAATACATTCTAGGACTTTATAAAAGATCACTTTTTATTTA
12	46414608	46414659	TGGCCATCTCTGACCTGT	TAGTGAGCCCAGAAGGGG	GACATCTACCTGCTCAACCTGGCCATCTCTGACCTGTTTTTCCTATTTACTGTCCCCTTCTGGGCTCACTATGCTGCCGCCCAGTGGGAC
13	46414768	46414820	TAGGTACCTGGCTGTCGT	TGACCGTCCTGGCTTTTA	TCATCCTCCTGACAATCGATAGGTACCTGGCTGTCGTCCATGCTACGTTTGCTTTAAAAGCCAGGACGGTCACCTTTGGGGTGGTGACAA
14	46415017	46415068	TGTCATGGTCATCTGCTA	TCGACACCGAAGCAGAGT	GGCTGGTCCTGCCGCTGCTTGTCATGGTCATCTGCTACTCGGGAGACCTAAAAACTCTGCTTCGGTGTCGAAATGAGAAGAAGAGGCACA
15	46420034	46420084	TGCCCCCGCGAGGCCACA	TCTGGAAGTTGAACACCC	GGCAAGCCTTGGGTCATACTGCCCCCGCGAGGCCACATTGGCAAGTCAGCAAGGGTGTTCAACTTCCAGACTTGGCCATGGAGAAGACAT

Table 2_4. HDR and NHEJ efficiency of re-TALEN/ssODN targeting *CCR5*

# targeting site	cell type	HDR	NHEJ	HDR detection limit based on Information analysis
1	PGP1-iPS	0.06%	0.26%	0.04%
2	PGP1-iPS	0.48%	0.07%	0.01%
3	PGP1-iPS	1.71%	0.41%	0.03%
4	PGP1-iPS	0.02%	0.04%	0.02%*
5	PGP1-iPS	0.52%	0.73%	0.00%
6	PGP1-iPS	0.06%	0.15%	0.00%
7	PGP1-iPS	0.01%	0.00%	0.01%*
8	PGP1-iPS	0.03%	0.06%	0.00%
9	PGP1-iPS	0.27%	1.25%	0.00%
10	PGP1-iPS	0.68%	0.27%	0.01%
11	PGP1-iPS	0.06%	0.03%	0.00%
12	PGP1-iPS	0.38%	1.47%	0.04%
13	PGP1-iPS	0.00%	0.13%	0.00%
14	PGP1-iPS	0.47%	0.37%	0.02%
15	PGP1-iPS	0.80%	0.14%	0.08%
1	K562	2.01%	0.89%	0.10%
2	K562	18.50%	12.82%	0.00%
3	K562	38.58%	67.93%	0.55%
4	K562	0.80%	2.47%	0.08%
5	K562	2.69%	64.74%	0.15%
6	K562	0.78%	1.34%	0.02%
7	K562	0.06%	0.77%	0.01%
8	K562	3.93%	12.34%	0.19%
9	K562	2.54%	2.96%	0.00%
10	K562	17.97%	6.08%	0.30%
11	K562	0.68%	0.24%	0.07%
12	K562	1.88%	0.46%	0.00%
13	K562	1.85%	0.26%	0.00%
14	K562	12.04%	0.83%	1.10%
15	K562	11.41%	4.14%	0.21%

* The group where HDR detection limit exceeds the real HDR detected

Table 2_5. CCR5 targeting site PCR primer sequences

# targeting in CCR5	name	primer sequence
1	site1-F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGATTTTGCAGTGTGCGTTACTCC
	site1-F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATCGTTTGCAGTGTGCGTTACTCC
	site1-F3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCTAATTTGCAGTGTGCGTTACTCC
	site1-F4	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGTCATTTGCAGTGTGCGTTACTCC
	site1-R	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCCAAGCAACTAAGTCACAGCA
2	Site2-F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGATATGAGGAAATGGAAGCTTG
	Site2-F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATCGATGAGGAAATGGAAGCTTG
	Site2-F3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCTAAATGAGGAAATGGAAGCTTG
	Site2-F4	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGTCAATGAGGAAATGGAAGCTTG
	Site2-R	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCATTAGGGTATTGGAGGA
3	site3-F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGATAATCCTCCCAACAACTCAT
	site3-F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATCGAATCCTCCCAACAACTCAT
	site3-F3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCTAAAATCCTCCCAACAACTCAT
	site3-F4	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGTCAAATCCTCCCAACAACTCAT
	site3_R	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCCCAATCCTACAGAGGCAG
4	site4-F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGATAAGCCAAAGCTTTTTATTC
	site4-F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATCGAAGCCAAAGCTTTTTATTC
	site4-F3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCTAAAAGCCAAAGCTTTTTATTC
	site4-F4	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGTCAAAGCCAAAGCTTTTTATTC
	site4_R	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGCCAAAGCTTTTTATTCT
5	site5-F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGATATCTTGTGGCTCGGGAGTAG
	site5-F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATCGATCTTGTGGCTCGGGAGTAG
	site5-R	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTTGGCAGGATTCTTCACTCCA
6	site6-F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGATCTATTTTGTTGCCCTTCAAA
	site6-F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATCGCTATTTTGTTGCCCTTCAAA
	site6-R	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTAACCTGAACTTGACCATATACT
7	site7-F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGATCAGCTGAGAGGTTACTTACC
	site7-F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATCGCAGCTGAGAGGTTACTTACC
	site7-R	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTAATGATTTAACTCCACCCTC
8	site8-F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGATACTCCACCCTCCTTCAAAAGA
	site8-F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATCGACTCCACCCTCCTTCAAAAGA
	site8-R	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTTGGTGTTTGCCAAATGTCT
9	site9_F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGATGGGCACATATTCAGAAGGCA
	site9_F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATCGGGGCACATATTCAGAAGGCA
	site9_R	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTAGTGAAAGACTTTAAAGGGAGCA
10	site10-F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGATCACAATTAAGAGTTGTCATA
	site10-F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATCGCACAATTAAGAGTTGTCATA
	site10-R	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCTCAGCTAGAGCAGCTGAAC
11	site11-F1	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGACACTTGATAATCCATC
	site11-F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATCGTCAATGTAGACATCTATGTAG
	site11-R	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGATTCAATGTAGACATCTATGTAG
12	site12-F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGATACTGCAAAAGGCTGAAGAGC
	site12-F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATCGACTGCAAAAGGCTGAAGAGC
	site12-F3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCTAAACTGCAAAAGGCTGAAGAGC
	site12-F4	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGTCAACTGCAAAAGGCTGAAGAGC
	site12-R	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGCCTATAAAATAGAGCCCTGTCAA
13	site13-F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGATCTCTATTTTATAGGCTTCTTC
	site13-F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATCGCTCTATTTTATAGGCTTCTTC
	site13-R	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTAGCCACCACCCAAGTGATC
14	site14-F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATCGTTCCAGACATTAAAGATAGTC
	site14-F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGATTTCCAGACATTAAAGATAGTC
	site14-R	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTAATCATGATGGTGAAGATAAG
15	site15-F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGATCCGGCAGAGACAAACATTAAA
	site15-F2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCGGCAGAGACAAACATTAAA
	site15-R	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTAGCTAGGAAGCCATGGCAAG
illumina adaptor	PE-PCR-F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACAcgac*g*c
	PE-PCR-R	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACc*g*c
Multiplex sequencing PCR primer		
3	site3-M-F	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTGCATAGTATGTGCTAGATGCTG
	site3-M-R	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTGATCTCTAAGAAGGCAAATGAGAC
illumina adaptor	Index-PCR	CAAGCAGAAGACGGCATACGAGATN$_1$N$_2$N$_3$N$_4$N$_5$N$_6$GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
	universal-PCR	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

*index-PCR primers are purchased from epicentre (ScriptSeq™ Index PCR Primers

Table 2_6. ssODN design for studying ssODN-mediated genome editing

Used in Figure	variation	ssODN name	ssODN sequence
Figure 7A	length of the ssODN	50	CCTCTGTAGTCAGACTGCCAAGCTTGAAACCTGTTATACCCTCAGCTTTCTCGTCTGGGTATTGGGCTGA
		70	CTACTGTCATTCAGCCCAATACCCAGACGAGAAAGCTGAGGGTATAACAGGTTTCAAGCTTGGCAGTCTGACTACAGAGGCCACTGGCTT
		90	CATGAAATGACTACTGTCATTCAGCCCAATACCCAGACGAGAAAGCTGAGGGTATAACAGGTTTCAAGCTTGGCAGTCTGACTACAGAGGCCACTGGCTTAGCCCCTGGG
		110	CCCAACAACTCATGAAATGACTACTGTCATTCAGCCCAATACCCAGACGAGAAAGCTGAGGGTATAACAGGTTTCAAGCTTGGCAGTCTGACTACAGAGGCCACTGGCTTAGC CCCTGGGTTAGTCTGCC
		130	ATAGAATCCTCCCAACAACTCATGAAATGACTACTGTCATTCAGCCCAATACCCAGACGAGAAAGCTGAGGGTATAACAGGTTTCAAGCTTGGCAGTCTGACTACAGAGGCCA CTGGCTTAGCCCCTGGGTTAGTCTGCCTCTGTAGGAT
		150	TGGATGCCTCATAGAATCCTCCCAACAACTCATGAAATGACTACTGTCATTCAGCCCAATACCCAGACGAGAAAGCTGAGGGTATAACAGGTTTCAAGCTTGGCAGTCTGACT ACAGAGGCCACTGGCTTAGCCCCTGGGTTAGTCTGCCTCTGTAGGATTGGGGGCACG
Figure 7B	Distance between middle of ssODN and DSB	-40	TGGATGCCTCATAGAATCCTCCCAACAACTCATGAAATGACTACTGTCATTCAGCCCAATACCCAGACGAGAAAGCTGAGGGTATAACAG
		-30	ATAGAATCCTCCCAACAACTCATGAAATGACTACTGTCATTCAGCCCAATACCCAGACGAGAAAGCTGAGGGTATAACAGGTTTCAAGCT
		-20	CCCAACAACTCATGAAATGACTACTGTCATTCAGCCCAATACCCAGACGAGAAAGCTGAGGGTATAACAGGTTTCAAGCTTGGCAGTCTG
		-10	CATGAAATGACTACTGTCATTCAGCCCAATACCCAGACGAGAAAGCTGAGGGTATAACAGGTTTCAAGCTTGGCAGTCTGACTACAGAGG
		0	CTACTGTCATTCAGCCCAATACCCAGACGAGAAAGCTGAGGGTATAACAGGTTTCAAGCTTGGCAGTCTGACTACAGAGGCCACTGGCTT
		10	TCAGCCCAATACCCAGACGAGAAAGCTGAGGGTATAACAGGTTTCAAGCTTGGCAGTCTGACTACAGAGGCCACTGGCTTAGCCCCTGGG
		20	ACCCAGACGAGAAAGCTGAGGGTATAACAGGTTTCAAGCTTGGCAGTCTGACTACAGAGGCCACTGGCTTAGCCCCTGGGTTAGTCTGCC
		30	GAAAGCTGAGGGTATAACAGGTTTCAAGCTTGGCAGTCTGACTACAGAGGCCACTGGCTTAGCCCCTGGGTTAGTCTGCCTCTGTAGGAT
		40	GGTATAACAGGTTTCAAGCTTGGCAGTCTGACTACAGAGGCCACTGGCTTAGCCCCTGGGTTAGTCTGCCTCTGTAGGATTGGGGGCACG
		-40	CTGTTATACCCTCAGCTTTCTCGTCTGGGTATTGGGCTGAATGACAGTAGTCATTTCATGAGTTGTTGGGAGGATTCTATGAGGCATCCA
		-30	AGCTTGAAACCTGTTATACCCTCAGCTTTCTCGTCTGGGTATTGGGCTGAATGACAGTAGTCATTTCATGAGTTGTTGGGAGGATTCTAT
		-20	CAGACTGCCAAGCTTGAAACCTGTTATACCCTCAGCTTTCTCGTCTGGGTATTGGGCTGAATGACAGTAGTCATTTCATGAGTTGTTGGG
		-10	CCTCTGTAGTCAGACTGCCAAGCTTGAAACCTGTTATACCCTCAGCTTTCTCGTCTGGGTATTGGGCTGAATGACAGTAGTCATTTCATG
		0	AAGCCAGTGGCCTCTGTAGTCAGACTGCCAAGCTTGAAACCTGTTATACCCTCAGCTTTCTCGTCTGGGTATTGGGCTGAATGACAGTAG
		10	CCCAGGGGCTAAGCCAGTGGCCTCTGTAGTCAGACTGCCAAGCTTGAAACCTGTTATACCCTCAGCTTTCTCGTCTGGGTATTGGGCTGA
		20	GGCAGACTAACCCAGGGGCTAAGCCAGTGGCCTCTGTAGTCAGACTGCCAAGCTTGAAACCTGTTATACCCTCAGCTTTCTCGTCTGGGT
		30	ATCCTACAGAGGCAGACTAACCCAGGGGCTAAGCCAGTGGCCTCTGTAGTCAGACTGCCAAGCTTGAAACCTGTTATACCCTCAGCTTTC
		40	CGTGCCCCCAATCCTACAGAGGCAGACTAACCCAGGGGCTAAGCCAGTGGCCTCTGTAGTCAGACTGCCAAGCTTGAAACCTGTTATACC
Figure 7C	Distance between the secondary mutation and DSB	90-*1	CTACTGTCATTCAGGGCAATACCCAGACGAGAAAGCTGAGGGTATAACAGGTTTCAAGCTTGGCAGTCTGACTACAGAGGCCACTGGCTT
		90-*2	CTACTGTCATTCAGCCCAATACCCTAACGAGAAAGCTGAGGGTATAACAGGTTTCAAGCTTGGCAGTCTGACTACAGAGGCCACTGGCTT
		90-*3	CTACTGTCATTCAGCCCAATACCCAGACGAGAAAAGTGAGGGTATAACAGGTTTCAAGCTTGGCAGTCTGACTACAGAGGCCACTGGCTT
		90M-0	CTACTGTCATTCAGCCCAATACCCAGACGAGAAAGCTGAGGGTATAACAGGTTTCAAGCTTGGCAGTCTGACTACAGAGGCCACTGGCTT
		90-*4	CTACTGTCATTCAGCCCAATACCCAGACGAGAAAGCTGAGGGTATAACAGGTTTGTAGCTTGGCAGTCTGACTACAGAGGCCACTGGCTT
		90-*5	CTACTGTCATTCAGCCCAATACCCAGACGAGAAAGCTGAGGGTATAACAGGTTTCAAGCTTGGCTCTCTGACTACAGAGGCCACTGGCTT
		90-*6	CTACTGTCATTCAGCCCAATACCCAGACGAGAAAGCTGAGGGTATAACAGGTTTCAAGCTTGGCAGTCTGACTAGTGAGGCCACTGGCTT
Figure 7D	distance between ssODN and the DSB	L670bp_90M	CACTTTATATTTCCCTGCTTAAACAGTCCCCCGAGGGTGGGTGCGGAAAAGGCTCTACACTTGTTATCATTCCCTCTCCACCACAGGCAT
		L570bp_90M	TTTGTATTTGGGTTTTTTTAAAACCTCCACTCTACAGTTAAGAATTCTAAGGCACAGAGCTTCAATAATTTGGTCAGAGCCAAGTAGCAG
		L480bp_90M	GGAGGTTAAACCCAGCAGCATGACTGCAGTTCTTAATCAATGCCCCTTGAATTGCACATATGGGATGAACTAGAACATTTTCTCGATGAT
		L394bp_90M	CTCGATGATTCGCTGTCCTTGTTATGATTATGTTACTGAGCTCTACTGTAGCACAGACATATGTCCCTATATGGGGCGGGGGTGGGGGTG
		L290bp_90M	GGTGTCTTGATCGCTGGGCTATTTCTATACTGTTCTGGCTTTTCGGAAGCAGTCATTTCTTTCTATTCTCCAAGCACCAGCAATTAGCTT
		L200bp_90M	GCTTCTAGTTTGCTGAAACTAATCTGCTATAGACAGAGACTCCGACGAACCAATTTTATTAGGATTTGATCAAATAAACTCTCTCTGACA
		L114bp_90M	GAAAGAGTAACTAAGAGTTTGATGTTTACTGAGTGCATAGTATGCACTAGATGCTGGCCGTGGATGCCTCATAGAATCCTCCCAACAACT
		L45bp_90M	GCTAGATGCTGGCCGTGGATGCCTCATAGAATCCTCCCAACAACCGATGAAATGACTACTGTCATTCAGCCCAATACCCAGACGAGAAAG
		R40bp_90M	ACAGGTTTCAAGCTTGGCAGTCTGACTACAGAGGCCACTGGCTTTACCCCTGGGTTAGTCTGCCTCTGTAGGATTGGGGGCACGTAATTT
		R100bp_90M	TTAGTCTGCCTCTGTAGGATTGGGGGCACGTAATTTTGCTGTTTAAGGTCTCATTTGCCTTCTTAGAGATCACAAGCCAAAGCTTTTTAT
		R200bp_90M	GGAAGCCCAGAGGGCATCTTGTGGCTCGGGAGTAGCTCTCTGCTACCTTCTCAGCTCTGCTGACAATACTTGAGATTTTCAGATGTCACC
		R261bp_90M	TCAGCTCTGCTGACAATACTTGAGATTTTCAGATGTCACCAACCAGCAAGAGAGCTTGATATGACTGTATATAGTATAGTCATAAAGAAC
		R322bp_90M	CATAAAGAACCTGAACTTGACCATATACTTATGTCATGTGGAAATCTTCTCATAGCTTCAGATAGATTATATCTGGAGTGAAGAATCCTG
		R375M_90M	GTGGAAAATTTCTCATAGCTTCAGATAGATTATATCTGGAGTGAGCAATCCTGCCACCTATGTATCTGGCATAGTGTGAGTCCTCATAAA
		R448bp_90M	GGTTTGAAGGGCAACAAAATAGTGAACAGAGTGAAAATCCCCACCTAGATCCTGGGTCCAGAAAAAGATGGGAAACCTGTTTAGCTCACC

sequence adaptor, 5 µl reaction products were then added to 20 µl of PCR mix containing 12.5 µl 2X KAPA HIFI Hotstart Readymix (KAPA Biosystems) and 200 nM primers carrying Illumina sequence adaptors. Reactions were incubated at 95°C for 5min followed by 25 cycles of 98°C, 20s; 65°C, 20s and 72°C, 20s. PCR products were purified by QIAquick PCR purification kit, mixed at roughly the same concentration, and sequenced with MiSeq Personal Sequencer. All the PCR primers can be found in the Table 2_5.

(3) Genome editing assessment system (GEAS)

We wrote a pipeline to analyze the genome engineering data. This pipeline is integrated in one single Unix module, which uses different tools such as R, BLAT, and FASTX Toolkit.

Barcode splitting: Groups of samples were pooled together and sequenced using MiSeq 150bp paired end (PE150) (Illumina Next Gen Sequencing), and later separated based on DNA barcodes using FASTX Toolkit.

- Quality filtering: We trimmed nucleotides with lower sequence quality (phred score<20). After trimming, reads shorter than 80 nucleotides were discarded.
- Mapping: We used BLAT to map the paired reads independently to the reference genome and we generated .psl files as output.
- Indel calling: We defined indels as the full length reads containing 2 blocks of matches in the alignment. Only reads following this pattern in both paired end reads were

- considered. As a quality control, we required the indel reads to possess minimal 70nt matching with the reference genome and both blocks to be at least 20 nt long. Size and position of indels were calculated by the positions of each block to the reference genome. Non-homologous end joining (NHEJ) has been estimated as the percentage of reads containing indels (see equation 1). The majority of NHEJ event have been detected at the targeting site vicinity.

- Homology directed recombination (HDR) efficiency: Pattern matching (grep) within a 12bp window centering over DSB was used to count specific signatures corresponding to reads containing the reference sequence, modifications of the reference sequence (2bp intended mismatches), and reads containing only 1bp mutation within the 2bp intended mismatches (see equation 1).

Equation 1. Estimation of NHEJ and HDR

A= reads identical to the reference: XXXXXABXXXXX

B= reads containing 2bp mismatch programed by ssODN: XXXXXabXXXXX

C= reads containing only 1 bp mutation in the target site: such as XXXXXaBXXXXX or XXXXXAbXXXXX

D = reads containing indels as described above

$$\text{NHEJ efficiency} = (100 \times \frac{D}{A + B + C + D})\%$$

$$\text{HDR efficiency} = (100 \times \frac{B}{A + B + C + D})\%$$

Statistical analysis of genome editing NGS data

(1) HDR specificity analysis

We used an exact binomial test to compute the probabilities of observing various numbers of sequence reads containing the 2bp mismatch. Based on the sequencing results of 10bp windows before and after the targeting site, we estimated the maximum base change rates of the two windows (P1 and P2). Using the null hypothesis that the changes of each of the two target bp were independent, we computed the expected probability of observing 2bp mismatch at the targeting site by chance as the product of these two probabilities (P1*P2). Given a dataset containing N numbers of total reads and n number of HDR reads, we calculated the p-value of the observed HDR efficiency.

(2) HDR sensitivity analysis

In our experimental design, the ssODN DNA donors contained a 2bp mismatch against the targeting genome, so that we expected co-presence of the base changes in the two target bp if the ssODN was incorporated into the targeting genome. Other non-intended observed sequence changes would not likely change at the same time. Thus, we predicted non-intended changes to be much less interdependent. Based on these assumptions, we used mutual information (MI) to measure the mutual dependence of simultaneous two base pair changes in all other pairs of positions, and we estimated the HDR detection limit as the smallest HDR where MI of the targeting 2bp site is higher than MI of all the other position pairs. For a given experiment, we first identified HDR reads with intended 2bp mismatch from the original fastq file and we simulated a set of fastq files with diluted HDR efficiencies by systematically removing different numbers of HDR reads from the original data set. Mutual information (MI) was computed between all pairs of positions within a 20bp window centered on the targeting site. In these

calculations, the mutual information of the base composition between any two positions is computed. Thus, unlike our HDR specificity measure above, this measure does not assess the tendency of position pairs to change to any particular pairs of target bases, only their tendency to change at the same time. We coded our analysis in R and MI was computed using the package infotheo.

(3) Correlations between genome editing efficiency and epigenetic state

We computed Pearson correlation coefficients to study possible associations between epigenetic parameters (DNase I HS or nucleosome occupancy) and genome engineering efficiencies (HDR, NHEJ). Dataset of epigenetic parameters for both cell types were downloaded from UCSC genome browser.

K562 cells HS: /gbdb/hg19/bbi/wgEncodeOpenChromDnaseK562SigV2.bigWig

hiPSCs DNase I HS: /gbdb/hg19/bbi/wgEncodeOpenChromDnaseIpsnihi7Sig.bigWig

K562 cells nucleosome occupancy: /gbdb/hg19/bbi/wgEncodeSydhNsomeK562Sig.bigWig

To compute P-values, we compared the observed correlation to a simulated distribution which was built by randomizing the position of the epigenetic parameter (N=100000). Observed correlations higher than the 95th percentile, or lower than the 5th percentile of the simulated distribution were considered as potential associations.

(4) Insertion composition analysis

We analyzed whether inserted fragments tend to contain ssODN sequence. We mapped the insertion sequence to the ssODN and counted the mapping occurrence at specific positions in

the ssODN. As the control, we generate a set of randomized DNA sequence with the same size profile of insertions and conducted the same mapping analysis (Figure S9).

Genotype screening of colonized hiPSCs

(1) FACS sorting of single-hiPSCs

Human iPS cells on feeder-free cultures were pre-treated with mTesr-1 media supplemented with SMC4 (5 uM thiazovivin,1 uM CHIR99021, 0.4 uM PD0325901, 2 uM SB431542) (*42*) for at least 2 hrs prior to FACS sorting. Cultures were dissociated using Accutase (Millipore) and resuspended in mTesr-1 media supplemented with SMC4 and the viability dye ToPro-3 (Invitrogen) at concentration of 1~2 X107 /mL. Live hiPS cells were single-cell sorted using a BD FACSAria II SORP UV (BD Biosciences) with 100um nozzle under sterile conditions into 96-well plates coated with irradiated CF-1 mouse embryonic fibroblasts (Global Stem). Each well contained hES cell medium (*59*) with 100 ng / ml recombinant human basic Fibroblast Growth Factor (bFGF) (Millipore) supplemented with SMC4 and 5 ug / ml fibronectin (Sigma). After sorting, plates were centrifuged at 70 x g for 3 min. Colony formation was seen 4 days post sorting, and the culture media was replaced with hES cell medium with SMC4. SMC4 can be removed from hES cell medium 8 days after sorting.

(2) Genotyping monoclonal hiPSCs

A few thousand cells were harvested 8 days after Fluorescence-activated cell sorting (FACS) and 0.1ul prepGEM tissue protease enzyme (ZyGEM) and 1ul prepGEM gold buffer (ZyGEM) were added to 8.9 µl of cells in the medium. The reactions were then added to 40 µl of PCR mix containing 35.5ml platinum 1.1X Supermix (Invitrogen), 250nM of each dNTP and

400nM primers. Reactions were incubated at 95°C for 3min followed by 30 cycles of 95°C, 20s; 65°C, 30s and 72°C, 20s. Products were Sanger sequenced using either one of the PCR primers (Table S5) and sequences were analyzed using DNASTAR (DNASTAR).

(3) Immunostaining

Cells were incubated in the KnockOut DMEM/F-12 medium at 37˚C for 60 minutes using the following antibody: Anti-SSEA-4 PE (Millipore) (1: 500 diluted); Tra-1-60 (BD Pharmingen) (1:100 diluted). After the incubation, cells were washed three times with KnockOut DMEM/F-12 and imaged on the Axio Observer Z.1 (ZIESS).

(4) Teratoma Formation and Analysis

Human iPSCs were harvested using collagenase type IV (Invitrogen) and resuspended into 200 µl of Matrigel and injected intramuscularly into the hind limbs of Rag2gamma knockout mice. Teratomas were isolated and fixed in formalin between 4 - 8 weeks after the injection. The teratomas were subsequently analyzed by hematoxylin and eosin staining.

Sequence used in the study

Sequence 2_1

re-TALE (16.5) sequence

```
CTAACCCCTGAACAGGTAGTCGCTATAGCTTCAAATATCGGGGGCAAGCAAGCACTTGAGACCGTTCAACGACTCCT
GCCAGTGCTCTGCCAAGCCCATGGATTGACTCCGGAGCAAGTCGTCGCGATCGCGAGCAACGGCGGGGGGAAGCAGG
CGCTGGAAACTGTTCAGAGACTGCTGCCTGTACTTTGTCAGGCGCATGGTCTCACCCCCGAACAGGTTGTCGCAATA
GCAAGTAATATAGGCGGTAAGCAAGCCCTAGAGACTGTGCAACGCCTGCTCCCCGTGCTGTGTCAGGCTCACGGTCT
GACACCTGAACAAGTTGTCGCGATAGCCAGTCACGACGGGGGAAAACAAGCTCTAGAAACGGTTCAAAGGTTGTTGC
CCGTTCTGTGCCAAGCACATGGGTTAACACCCGAACAAGTAGTAGCGATAGCGTCAAATAACGGGGGTAAACAGGCT
TTGGAGACGGTACAGCGGTTATTGCCGGTCCTCTGCCAGGCCCACGGACTTACGCCAGAACAGGTGGTTGCAATTGC
CTCCAACATCGGCGGGAAACAAGCGTTGGAAACTGTGCAGAGACTCCTTCCTGTTTTGTGTCAAGCCCACGGCTTGA
```

CGCCTGAGCAGGTTGTGGCCATCGCTAGCCACGACGGAGGGAAGCAGGCTCTTGAAACCGTACAGCGACTTCTCCCA
GTTTTGTGCCAAGCTCACGGGCTAACCCCCGAGCAAGTAGTTGCCATAGCAAGCAACGGAGGAGGAAAACAGGCATT
AGAAACAGTTCAGCGCTTGCTCCCGGTACTCTGTCAGGCACACGGTCTAACTCCGGAACAGGTCGTAGCCATTGCTT
CCCATGATGGCGGCAAACAGGCGCTAGAGACAGTCCAGAGGCTCTTGCCTGTGTTATGCCAGGCACATGGCCTCACC
CCGGAGCAGGTCGTTGCCATCGCCAGTAATATCGGCGGAAAGCAAGCTCTCGAAACAGTACAACGGCTGTTGCCAGT
CCTATGTCAAGCTCATGGACTGACGCCCGAGCAGGTAGTGGCAATCGCATCTCACGATGGAGGTAAACAAGCACTCG
AGACTGTCCAAAGATTGTTACCCGTACTATGCCAAGCGCATGGTTTAACCCCAGAGCAAGTTGTGGCTATTGCATCT
AACGGCGGTGGCAAACAAGCCTTGGAGACAGTGCAACGATTACTGCCTGTCTTATGTCAGGCCCATGGCCTTACTCC
TGAGCAAGTCGTAGCTATCGCCAGCAACATAGGTGGGAAACAGGCCCTGGAAACCGTACAACGTCTCCTCCCAGTAC
TTTGTCAAGCACACGGGTTGACACCGGAACAAGTGGTGGCGATTGCGTCCAACGGCGGAGGCAAGCAGGCACTGGAG
ACCGTCCAACGGCTTCTTCCGGTTCTTTGCCAGGCTCATGGGCTCACGCCAGAGCAGGTGGTAGCAATAGCGTCGAA
CATCGGTGGTAAGCAAGCGCTTGAAACGGTCCAGCGTCTTCTGCCGGTGTTGTGCCAGGCGCACGGACTCACACCAG
AACAAGTGGTTGCTATTGCTAGTAACAACGGTGGAAAGCAGGCCCTCGAGACGGTGCAGAGGTTACTTCCCGTCCTC
TGTCAAGCGCACGGCCTCACTCCAGAGCAAGTGGTTGCGATCGCTTCAAACAATGGTGGAAGACCTGCCCTGGAA

Sequence 2_2

re-TALEN-backbone sequence

(purple: re-TALE-N; red: SapI site; green: 0.5 monomer; blue: re-TALEN-C; orange: Fok I)

ATGTCGCGGACCCGGCTCCCTTCCCCACCCGCACCCAGCCCAGCGTTTTCGGCCGACTCGTTCTCAGACCTGCTTAG
GCAGTTCGACCCCTCACTGTTTAACACATCGTTGTTCGACTCCCTTCCTCCGTTTGGGGCGCACCATACGGAGGCGG
CCACCGGGGAGTGGGATGAGGTGCAGTCGGGATTGAGAGCTGCGGATGCACCACCCCCAACCATGCGGGTGGCCGTC
ACCGCTGCCCGACCGCCGAGGGCGAAGCCCGCACCAAGGCGGAGGGCAGCGCAACCGTCCGACGCAAGCCCCGCAGC
GCAAGTAGATTTGAGAACTTTGGGATATTCACAGCAGCAGCAGGAAAAGATCAAGCCCAAAGTGAGGTCGACAGTCG
CGCAGCATCACGAAGCGCTGGTGGGTCATGGGTTTACACATGCCCACATCGTAGCCTTGTCGCAGCACCCTGCAGCC
CTTGGCACGGTCGCCGTCAAGTACCAGGACATGATTGCGGCGTTGCCGGAAGCCACACATGAGGCGATCGTCGGTGT
GGGGAAACAGTGGAGCGGAGCCCGAGCGCTTGAGGCCCTGTTGACGGTCGCGGGAGAGCTGAGAGGGCCTCCCCTTC
AGCTGGACACGGGCCAGTTGCTGAAGATCGCGAAGCGGGGAGGAGTCACGGCGGTCGAGGCGGTGCACGCGTGGCGC
AATGCGCTCACGGGAGCACCCCTCAACAGTTCACGCTGACAGAGACCGCGGCCGCATTAGGCACCCCAGGCTTTACA
CTTTATGCTTCCGGCTCGTATAATGTGTGGATTTTGAGTTAGGATCCGTCGAGATTTTCAGGAGCTAAGGAAGCTAA
AATGGAGAAAAAAATCACTGGATATACCACCGTTGATATATCCCAATGGCATCGTAAAGAACATTTTGAGGCATTTC
AGTCAGTTGCTCAATGTACCTATAACCAGACCGTTCAGCTGGATATTACGGCCTTTTTAAAGACCGTAAAGAAAAAT
AAGCACAAGTTTTATCCGGCCTTTATTCACATTCTTGCCCGCCTGATGAATGCTCATCCGGAATTCCGTATGGCAAT
GAAAGACGGTGAGCTGGTGATATGGGATAGTGTTCACCCTTGTTACACCGTTTTCCATGAGCAAACTGAAACGTTTT
CATCGCTCTGGAGTGAATACCACGACGATTTCCGGCAGTTTCTACACATATATTCGCAAGATGTGGCGTGTTACGGT
GAAAACCTGGCCTATTTCCCTAAAGGGTTTATTGAGAATATGTTTTTCGTCTCAGCCAATCCCTGGGTGAGTTTCAC
CAGTTTTGATTTAAACGTGGCCAATATGGACAACTTCTTCGCCCCCGTTTTCACCATGGGCAAATATTATACGCAAG
GCGACAAGGTGCTGATGCCGCTGGCGATTCAGGTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCTT
AATGAATTACAACAGTACTGCGATGAGTGGCAGGGCGGGGCGTAAAGATCTGGATCCGGCTTACTAAAAGCCAGATA
ACAGTATGCGTATTTGCGCGCTGATTTTTGCGGTATAAGAATATATACTGATATGTATACCCGAAGTATGTCAAAAA
GAGGTATGCTATGAAGCAGCGTATTACAGTGACAGTTGACAGCGACAGCTATCAGTTGCTCAAGGCATATATGATGT
CAATATCTCCGGTCTGGTAAGCACAACCATGCAGAATGAAGCCCGTCGTCTGCGTGCCGAACGCTGGAAAGCGGAAA
ATCAGGAAGGGATGGCTGAGGTCGCCCGGTTTATTGAAATGAACGGCTCTTTTGCTGACGAGAACAGGGGCTGGTGA
AATGCAGTTTAAGGTTTACACCTATAAAAGAGAGAGCCGTTATCGTCTGTTTGTGGATGTACAGAGTGATATTATTG
ACACGCCCGGGCGACGGATGGTGATCCCCCTGGCCAGTGCACGTCTGCTGTCAGATAAAGTCTCCCGTGAACTTTAC
CCGGTGGTGCATATCGGGGATGAAAGCTGGCGCATGATGACCACCGATATGGCCAGTGTGCCGGTCTCCGTTATCGG
GGAAGAAGTGGCTGATCTCAGCCACCGCGAAAATGACATCAAAAACGCCATTAACCTGATGTTCTGGGGAATATAAA

TGTCAGGCTCCCTTATACACAGCCAGTCTGCAGGTCGACGGTCTCGCTCTTCGAAGGTTACTTCCCGTCCTCTGTCA
AGCGCACGGCCTCACTCCAGAGCAAGTGGTTGCGATCGCTTCAAACAACGGTGGAAGACCTGCCCTGGAATCAATCG
TGGCCCAGCTTTCGAGGCCGGACCCCGCGCTGGCCGCACTCACTAATGATCATCTTGTAGCGCTGGCCTGCCTCGGC
GGACGACCCGCCTTGGATGCGGTGAAGAAGGGGCTCCCGCACGCGCCTGCATTGATTAAGCGGACCAACAGAAGGAT
TCCCGAGAGGACATCACATCGAGTGGCAGGTTCCCAACTCGTGAAGAGTGAACTTGAGGAGAAAAAGTCGGAGCTGC
GGCACAAATTGAAATACGTACCGCATGAATACATCGAACTTATCGAAATTGCTAGGAACTCGACTCAAGACAGAATC
CTTGAGATGAAGGTAATGGAGTTCTTTATGAAGGTTTATGGATACCGAGGGAAGCATCTCGGTGGATCACGAAAACC
CGACGGAGCAATCTATACGGTGGGGAGCCCGATTGATTACGGAGTGATCGTCGACACGAAAGCCTACAGCGGTGGGT
ACAATCTTCCCATCGGGCAGGCAGATGAGATGCAACGTTATGTCGAAGAAAATCAGACCAGGAACAAACACATCAAT
CCAAATGAGTGGTGGAAAGTGTATCCTTCATCAGTGACCGAGTTTAAGTTTTTGTTTGTCTCTGGGCATTTCAAAGG
CAACTATAAGGCCCAGCTCACACGGTTGAATCACATTACGAACTGCAATGGTGCGGTTTTGTCCGTAGAGGAACTGC
TCATTGGTGGAGAAATGATCAAAGCGGGAACTCTGACACTGGAAGAAGTCAGACGCAAGTTTAACAATGGCGAGATC
AATTTCCGC

re-TALE-TF backbone sequence

(purple: re-TALE-N; red: SapI site; green: 0.5 monomer; blue: re-TALEN-C; orange: NLS-VP64; 2A-GFP is highlighted in green)

ATGTCGCGGACCCGGCTCCCTTCCCCACCCGCACCCAGCCCAGCGTTTTCGGCCGACTCGTTCTCAGACCTGCTTAG
GCAGTTCGACCCCTCACTGTTTAACACATCGTTGTTCGACTCCCTTCCTCCGTTTGGGGCGCACCATACGGAGGCGG
CCACCGGGGAGTGGGATGAGGTGCAGTCGGGATTGAGAGCTGCGGATGCACCACCCCCAACCATGCGGGTGGCCGTC
ACCGCTGCCCGACCGCCGAGGGCGAAGCCCGCACCAAGGCGGAGGGCAGCGCAACCGTCCGACGCAAGCCCCGCAGC
GCAAGTAGATTTGAGAACTTTGGGATATTCACAGCAGCAGCAGGAAAAGATCAAGCCCAAAGTGAGGTCGACAGTCG
CGCAGCATCACGAAGCGCTGGTGGGTCATGGGTTTACACATGCCCACATCGTAGCCTTGTCGCAGCACCCTGCAGCC
CTTGGCACGGTCGCCGTCAAGTACCAGGACATGATTGCGGCGTTGCCGGAAGCCACACATGAGGCGATCGTCGGTGT
GGGGAAACAGTGGAGCGGAGCCCGAGCGCTTGAGGCCCTGTTGACGGTCGCGGGAGAGCTGAGAGGGCCTCCCCTTC
AGCTGGACACGGGCCAGTTGCTGAAGATCGCGAAGCGGGGAGGAGTCACGGCGGTCGAGGCGGTGCACGCGTGGCGC
AATGCGCTCACGGGAGCACCCCTCAACAGTTCACGCTGACAGAGACCGCGGCCGCATTAGGCACCCCAGGCTTTACA
CTTTATGCTTCCGGCTCGTATAATGTGTGGATTTTGAGTTAGGATCCGTCGAGATTTTCAGGAGCTAAGGAAGCTAA
AATGGAGAAAAAAATCACTGGATATACCACCGTTGATATATCCCAATGGCATCGTAAAGAACATTTTGAGGCATTTC
AGTCAGTTGCTCAATGTACCTATAACCAGACCGTTCAGCTGGATATTACGGCCTTTTTAAAGACCGTAAAGAAAAAT
AAGCACAAGTTTTATCCGGCCTTTATTCACATTCTTGCCCGCCTGATGAATGCTCATCCGGAATTCCGTATGGCAAT
GAAAGACGGTGAGCTGGTGATATGGGATAGTGTTCACCCTTGTTACACCGTTTTCCATGAGCAAACTGAAACGTTTT
CATCGCTCTGGAGTGAATACCACGACGATTTCCGGCAGTTTCTACACATATATTCGCAAGATGTGGCGTGTTACGGT
GAAAACCTGGCCTATTTCCCTAAAGGGTTTATTGAGAATATGTTTTTCGTCTCAGCCAATCCCTGGGTGAGTTTCAC
CAGTTTTGATTTAAACGTGGCCAATATGGACAACTTCTTCGCCCCCGTTTTCACCATGGGCAAATATTATACGCAAG
GCGACAAGGTGCTGATGCCGCTGGCGATTCAGGTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCTT
AATGAATTACAACAGTACTGCGATGAGTGGCAGGGCGGGGCGTAAAGATCTGGATCCGGCTTACTAAAAGCCAGATA
ACAGTATGCGTATTTGCGCGCTGATTTTTGCGGTATAAGAATATATACTGATATGTATACCCGAAGTATGTCAAAAA
GAGGTATGCTATGAAGCAGCGTATTACAGTGACAGTTGACAGCGACAGCTATCAGTTGCTCAAGGCATATATGATGT
CAATATCTCCGGTCTGGTAAGCACAACCATGCAGAATGAAGCCCGTCGTCTGCGTGCCGAACGCTGGAAAGCGGAAA
ATCAGGAAGGGATGGCTGAGGTCGCCCGGTTTATTGAAATGAACGGCTCTTTTGCTGACGAGAACAGGGGCTGGTGA
AATGCAGTTTAAGGTTTACACCTATAAAAGAGAGAGCCGTTATCGTCTGTTTGTGGATGTACAGAGTGATATTATTG
ACACGCCCGGGCGACGGATGGTGATCCCCCTGGCCAGTGCACGTCTGCTGTCAGATAAAGTCTCCCGTGAACTTTAC
CCGGTGGTGCATATCGGGGATGAAAGCTGGCGCATGATGACCACCGATATGGCCAGTGTGCCGGTCTCCGTTATCGG
GGAAGAAGTGGCTGATCTCAGCCACCGCGAAAATGACATCAAAAACGCCATTAACCTGATGTTCTGGGGAATATAAA
TGTCAGGCTCCCTTATACACAGCCAGTCTGCAGGTCGACGGTCTCGCTCTTCGAAGGTTACTTCCCGTCCTCTGTCA
AGCGCACGGCCTCACTCCAGAGCAAGTGGTTGCGATCGCTTCAAACAACGGTGGAAGACCTGCCCTGGAATCAATCG
TGGCCCAGCTTTCGAGGCCGGACCCCGCGCTGGCCGCACTCACTAATGATCATCTTGTAGCGCTGGCCTGCCTCGGC

GGACGACCCGCCTTGGATGCGGTGAAGAAGGGGCTCCCGCACGCGCCTGCATTGATTAAGCGGACCAACAGAAGGAT
TCCCGAGAGGACATAGCCCCAAGAAGAAGAGAAAGGTGGAGGCCAGCGGTTCCGGACGGGCTGACGCATTGGACGAT
TTTGATCTGGATATGCTGGGAAGTGACGCCCTCGATGATTTTGACCTTGACATGCTTGGTTCGGATGCCCTTGATGA
CTTTGACCTCGACATGCTCGGCAGTGACGCCCTTGATGATTTCGACCTGGACATGCTGATTAACTCTAGAGGCAGTG
GAGAGGGCAGAGGAAGTCTGCTAACATGCGGTGACGTCGAGGAGAATCCTGGCCCAGTGAGCAAGGGCGAGGAGCTG
TTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGG
CGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCA
CCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTC
AAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGC
CGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACA
TCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATC
AAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCC
CATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACG
AGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAG

References

1. S. O'Rahilly, Human genetics illuminates the paths to metabolic disease., *Nature* **462**, 307–14 (2009).

2. S. T. Sherry *et al.*, dbSNP: the NCBI database of genetic variation., *Nucleic acids research* **29**, 308–11 (2001).

3. T. Barrett *et al.*, BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata., *Nucleic acids research* **40**, D57–63 (2012).

4. H. Zhu, M. W. Lensch, P. Cahan, G. Q. Daley, Investigating monogenic and complex diseases with pluripotent stem cells., *Nature reviews. Genetics* **12**, 266–75 (2011).

5. M. H. Porteus, J. P. Connelly, S. M. Pruett, A Look to Future Directions in Gene Therapy Research for Monogenic Diseases, **2** (2006), doi:10.1371/journal.pgen.0020133.

6. J. Zou, P. Mali, X. Huang, S. N. Dowey, L. Cheng, Site-specific gene correction of a point mutation in human iPS cells derived from an adult patient with sickle cell disease., *Blood* **118**, 4599–608 (2011).

7. K. Yusa *et al.*, Targeted gene correction of α1-antitrypsin deficiency in induced pluripotent stem cells., *Nature* **478**, 391–4 (2011).

8. T. Mashimo *et al.*, E. A. A. Nollen, Ed. Generation of knockout rats with X-linked severe combined immunodeficiency (X-SCID) using zinc-finger nucleases., *PloS one* **5**, e8870 (2010).

9. F. Herrmann *et al.*, p53 Gene repair with zinc finger nucleases optimised by yeast 1-hybrid and validated by Solexa sequencing., *PloS one* **6**, e20913 (2011).

10. E. E. Perez *et al.*, Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases., *Nature biotechnology* **26**, 808–16 (2008).

11. N. Holt *et al.*, Human hematopoietic stem/progenitor cells modified by zinc-finger nucleases targeted to CCR5 control HIV-1 in vivo., *Nature biotechnology* **28**, 839–47 (2010).

12. D. Carroll, Genome engineering with zinc-finger nucleases., *Genetics* **188**, 773–82 (2011).

13. A. J. Wood *et al.*, Targeted genome editing across species using ZFNs and TALENs., *Science (New York, N.Y.)* **333**, 307 (2011).

14. P. Perez-Pinera, D. G. Ousterout, C. A. Gersbach, Advances in targeted genome editing., *Current opinion in chemical biology* **16**, 268–77 (2012).

15. M. L. Maeder *et al.*, Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification., *Molecular cell* **31**, 294–301 (2008).

16. F. D. Urnov *et al.*, Highly efficient endogenous human gene correction using designed zinc-finger nucleases., *Nature* **435**, 646–51 (2005).

17. L. Cade *et al.*, Highly efficient generation of heritable zebrafish gene mutations using homo- and heterodimeric TALENs, *Nucleic Acids Research* , 1–10 (2012).

18. D. Hockemeyer *et al.*, Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases., *Nature biotechnology* **27**, 851–7 (2009).

19. D. Hockemeyer *et al.*, Genetic engineering of human pluripotent cells using TALE nucleases., *Nature biotechnology* **29**, 731–4 (2011).

20. J. Zou *et al.*, Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells., *Cell stem cell* **5**, 97–110 (2009).

21. J. D. Sander *et al.*, Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA)., *Nature methods* **8**, 67–9 (2011).

22. J. Boch *et al.*, Breaking the code of DNA binding specificity of TAL-type III effectors., *Science (New York, N.Y.)* **326**, 1509–12 (2009).

23. C. Mussolino *et al.*, A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity., *Nucleic acids research* **39**, 9283–93 (2011).

24. D. Reyon *et al.*, FLASH assembly of TALENs for high-throughput genome editing, *Nature Biotechnology* **30**, 460–465 (2012).

25. A. W. Briggs *et al.*, Iterative capped assembly: rapid and scalable synthesis of repeat-module DNA such as TAL effectors from individual monomers., *Nucleic acids research* , 1–10 (2012).

26. F. Zhang *et al.*, LETTErs Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription, **29**, 149–154 (2011).

27. T. Cermak *et al.*, Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting., *Nucleic acids research* **39**, e82 (2011).

28. A. Lombardo *et al.*, Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery., *Nature biotechnology* **25**, 1298–306 (2007).

29. F. Chen *et al.*, High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases., *Nature methods* **8**, 753–5 (2011).

30. F. Soldner *et al.*, Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations., *Cell* **146**, 318–31 (2011).

31. S. Radecke, F. Radecke, T. Cathomen, K. Schwarz, Zinc-finger nuclease-induced gene repair with oligodeoxynucleotides: wanted and unwanted target locus modifications., *Molecular therapy : the journal of the American Society of Gene Therapy* **18**, 743–53 (2010).

32. F. Radecke *et al.*, Targeted chromosomal gene modification in human cells by single-stranded oligodeoxynucleotides in the presence of a DNA double-strand break., *Molecular therapy : the journal of the American Society of Gene Therapy* **14**, 798–808 (2006).

33. V. M. Bedell *et al.*, In vivo genome editing using a high-efficiency TALEN system., *Nature* **490**, 114–118 (2012).

34. J. Tian, K. Ma, I. Saaem, Advancing high-throughput gene synthesis technology., *Molecular bioSystems* **5**, 714–22 (2009).

35. D. G. Gibson *et al.*, Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome., *Science (New York, N.Y.)* **319**, 1215–20 (2008).

36. M. Li, S. Elledge, Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC, *Nature methods* **4**, 251–256 (2007).

37. J. Quan, J. Tian, Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries., *Nature protocols* **6**, 242–51 (2011).

38. C. E. Thomas, A. Ehrhardt, M. a Kay, Progress and problems with the use of viral vectors for gene therapy., *Nature reviews. Genetics* **4**, 346–58 (2003).

39. N. Holt *et al.*, Human hematopoietic stem/progenitor cells modified by zinc-finger nucleases targeted to CCR5 control HIV-1 in vivo., *Nature biotechnology* **28**, 839–47 (2010).

40. L. S. Symington, J. Gautier, Double-strand break end resection and repair pathway choice., *Annual review of genetics* **45**, 247–71 (2011).

41. S. Bannwarth, V. Procaccio, V. Paquis-Flucklinger, Surveyor Nuclease: a new strategy for a rapid identification of heteroplasmic mitochondrial DNA mutations in patients with respiratory chain defects., *Human mutation* **25**, 575–82 (2005).

42. B. Valamehr *et al.*, A novel platform to enable the high-throughput derivation and characterization of feeder-free human iPSCs., *Scientific reports* **2**, 213 (2012).

43. A. P. Boyle *et al.*, High-resolution mapping and characterization of open chromatin across the genome., *Cell* **132**, 311–22 (2008).

44. J. Streubel, C. Blücher, A. Landgraf, J. Boch, TAL effector RVD specificities and efficiencies., *Nature biotechnology* **30**, 593–5 (2012).

45. A. Valouev *et al.*, A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning., *Genome research* **18**, 1051–63 (2008).

46. S. J. Orlando *et al.*, Zinc-finger nuclease-driven targeted integration into mammalian genomes using donors with limited chromosomal homology., *Nucleic acids research* **38**, e152 (2010).

47. Z. Wang, Z.-J. Zhou, D.-P. Liu, J.-D. Huang, Double-stranded break can be repaired by single-stranded oligonucleotides via the ATM/ATR pathway in mammalian cells., *Oligonucleotides* **18**, 21–32 (2008).

48. X. Rios *et al.*, Stable gene targeting in human cells using single-strand oligonucleotides with modified bases., *PloS one* **7**, e36697 (2012).

49. B. Elliott *et al.*, Gene Conversion Tracts from Double-Strand Break Repair in Mammalian Cells Gene Conversion Tracts from Double-Strand Break Repair in Mammalian Cells, **18** (1998).

50. M. Jinek *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity., *Science (New York, N.Y.)* **337**, 816–21 (2012).

51. N. E. Sanjana *et al.*, A transcription activator-like effector toolbox for genome engineering., *Nature protocols* **7**, 171–92 (2012).

52. L. Cong, R. Zhou, Y. Kuo, M. Cunniff, F. Zhang, Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains, *Nature Communications* **3**, 968 (2012).

53. H. Fung, D. M. Weinstock, Repair at single targeted DNA double-strand breaks in pluripotent and differentiated human cells., *PloS one* **6**, e20514 (2011).

54. S. S. Ajay, S. C. J. Parker, H. O. Abaan, K. V. F. Fajardo, E. H. Margulies, Accurate and comprehensive sequencing of personal genomes., *Genome research* **21**, 1498–505 (2011).

55. H. Y. K. Lam *et al.*, Performance comparison of whole-genome sequencing platforms., *Nature biotechnology* **30**, 78–82 (2012).

56. R. Gabriel *et al.*, An unbiased genome-wide analysis of zinc-finger nuclease specificity., *Nature biotechnology* **29**, 816–23 (2011).

57. V. Pattanayak, C. L. Ramirez, J. K. Joung, D. R. Liu, Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection., *Nature methods* **8**, 765–70 (2011).

58. D. G. Gibson *et al.*, Enzymatic assembly of DNA molecules up to several hundred kilobases, **6**, 12–16 (2009).

59. I.-H. Park, P. H. Lerou, R. Zhao, H. Huo, G. Q. Daley, Generation of human-induced pluripotent stem cells., *Nature protocols* **3**, 1180–6 (2008).

Chapter 3

RNA-guided human genome engineering via Cas9

Prashant Mali[1,*], Luhan Yang[1,*], Kevin M. Esvelt[2], John Aach[1], Marc Guell[1], James E. DiCarlo[3], Julie Norville[1], P Benjamin Stranges[1], Mark Moosburner[1], Sriram Kosuri[1,2], George M. Church[1,2]

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA.
[2]Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, Massachusetts, USA.
[3]Department of Biomedical Engineering, Boston University, Boston, MA, USA,
[*]These authors contributed equally to this work

Acknowledgements

This work was supported by NIH grant P50 HG005550.

Author contribution

P.M., L.Y., and G.M.C. conceived the study jointly with K.M.E., M.G. and J.A. designed and performed the experiments with assistance from P.B.S, M.M., S.K, J.E.D. and J.N.; P.M. wrote the manuscripts with the help of all the other co-authors. G.M.C. supervised all aspects of the study. In particular, I devised and conducted the multiplexible targeting experiment. I conducted the targeting experiment and data analysis for 293, K562 and hiPSC, compared the activity and editing pattern of Cas9-nuclease and nickase.
In addition, I compared the genome targeting efficiency of reTALE and Cas9 system.

This chapter contains work from the manuscripts "RNA-guided human genome engineering via CAS9", published in *Science*, February 15, 2013, Vol. 339 no. 6121 pp. 823-826; and "CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering", published in *Nature Biotechnology*, August 1, 2013. The text and figures were modified to fit the format of this dissertation. The text and figures were modified to fit the format of this dissertation.

Summary

Bacteria and archaea have evolved adaptive immune defenses, termed clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) systems, that use short RNA to direct degradation of foreign nucleic acids. Here, we engineer the type II bacterial CRISPR system to function with custom guide RNA (gRNA) in human cells and build a genome-wide resource of ~190 K unique gRNAs targeting ~40.5% of human exons. In addition, we investigated off-target binding by Cas9-gRNA complexes and compared them with TAL effector (TALE) proteins and demonstrate methods to mitigate off-target phenomena by engineering a requirement for cooperatively through offset nicking for genome editing. Our results establish an RNA-guided editing tool for facile, robust, and multiplexable human genome engineering.

Introduction

Bacteria and archaea have evolved adaptive immune defenses termed clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) systems that use short RNA to direct degradation of foreign nucleic acids. CRISPR defense involves acquisition and integration of new targeting “spacers” from invading virus or plasmid DNA into the CRISPR locus, expression and processing of short guiding CRISPR RNAs (crRNAs) consisting of spacer-repeat units, and cleavage of nucleic acids (most commonly DNA) complementary to the spacer.

Three classes of CRISPR systems have been described thus far (Type I, II and III). Here we focus on the Type II CRISPR system, which utilizes a single effector enzyme, Cas9, to cleave dsDNA, whereas Type I and Type III systems require multiple distinct effectors acting as a

complex (for a detailed review of CRISPR classification, see reference (*1*)). As a consequence, Type II systems are more likely to function in alternative contexts such as eukaryotic cells. The Type II effector system consists of a long pre-crRNA transcribed from the spacer-containing CRISPR locus, the multifunctional Cas9 protein, and a tracrRNA important for gRNA processing. The tracrRNAs hybridize to the repeat regions separating the spacers of the pre-crRNA, initiating dsRNA cleavage by endogenous RNase III, which is followed by a second cleavage event within each spacer by Cas9, producing mature crRNAs that remain associated with the tracrRNA and Cas9. Jinek et al. demonstrated that a tracrRNA-crRNA fusion, termed a guide RNA (gRNA) in this work, is functional in vitro, obviating the need for RNase III and the crRNA processing in general (*2*).

Type II CRISPR interference is a result of Cas9 unwinding the DNA duplex and searching for sequences matching the crRNA to cleave. Target recognition occurs upon detection of complementarity between a "protospacer" sequence in the target DNA and the remaining spacer sequence in the crRNA. Importantly, Cas9 cuts the DNA only if a correct protospacer-adjacent motif (PAM) is also present at the 3' end. Different Type II systems have differing PAM requirements. The S. pyogenes system utilized in this work requires an NGG sequence, where N can be any nucleotide. S. thermophilus Type II systems require NGGNG (*3*) and NNAGAAW (*4*), respectively, while different S. mutans systems tolerate NGG or NAAR (*5*). Bioinformatic analyses have generated extensive databases of CRISPR loci in a variety of bacteria that may serve to identify new PAMs and expand the set of CRISPR-targetable sequences (*6*, *7*). In S. *thermophilus*, Cas9 generates a blunt-ended double-stranded break 3bp prior to the 3' end of the protospacer (*8*), a process mediated by two catalytic domains in the Cas9 protein: an HNH domain that cleaves the complementary strand of the DNA and a RuvC-

like domain that cleaves the non-complementary strand. While the S. pyogenes system has not been characterized to the same level of precision, DSB formation also occurs towards the 3' end of the protospacer. If one of the two nuclease domains is inactivated, Cas9 will function as a nickase in vitro (*2*) and in human cells.

As a genome engineering tool, the specificity of gRNA-directed Cas9 cleavage will be of the utmost importance. Significant off-target activity could cause unwanted double-strand breaks at other regions of the genome, resulting in toxicity and possibly oncogenesis in gene therapy applications. The S. pyogenes system tolerates mismatches in the first 6 bases out of the 20bp mature spacer sequence in vitro. However, it is entirely possible that greater stringency is required in vivo given the low toxicity we observed in human cell lines, as potential off-target sites matching (last 14 bp) NGG exist within the human reference genome for our gRNAs. Mismatches towards the 3' end of the spacer, known as the "seed sequence" (*9*), are less well tolerated. Jinek et al. found that single mismatches in the PAM at positions -3 through -7 abolished interference in vitro, though a mismatch at position -10 did not (*2*). In S. thermophilus, single mutations in the PAM or at positions -1, -3 through -5, and -7 through -8 abolished interference. When transplanted into E. coli, the S. thermophilus system did not tolerate single mutations in the PAM or in positions -3, -6, or -8. As a caveat, Garneau et al. found that spacers acquired from plasmid DNA tolerated greater degeneracy in both the PAM and seed sequence while sufficing to block plasmid acquisition in S. thermophiles (*10*); however, similar degeneracy was not sufficient to block phage infection (*11*), emphasizing the importance of the assay utilized. Taken together, these results point towards the urgent need to assay specificity in the context of interest.

Here, we sought to test whether we can engineer Cas9 system in human setting to conduct sequence specific genome engineering.

Results

Design of Cas9-gRNA system in human cells

First, we engineer the protein and RNA components of this bacterial type II CRISPR system in human cells. We began by synthesizing a human codon optimized version of the Cas9 protein bearing a C-terminal SV40 nuclear localization signal and cloning it into a mammalian expres- sion system (Figure 3_1 A, Figure 3_2 A).To direct Cas9 to cleave sequences of interest, we expressed crRNA-tracrRNA fusion transcripts, hereafter referred to as guide RNAs (gRNAs), from the human U6 polymerase III promoter. Directly transcribing gRNAs allowed us to avoid reconstituting the RNA-processing machinery used by bacterial CRISPR systems (Figure 3_1 A and Figure 3_2 B) (*2*, *12*), Constrained only by U6 transcription initiating with G and the requirement for the PAM (protospacer-adjacent motif) sequence –NGG following the 20–base pair (bp) crRNA target, our highly versatile approach can, in principle, target any genomic site of the form GN20GG (Figure 3_2C).

Functional test of Cas9-gRNA system in human cells

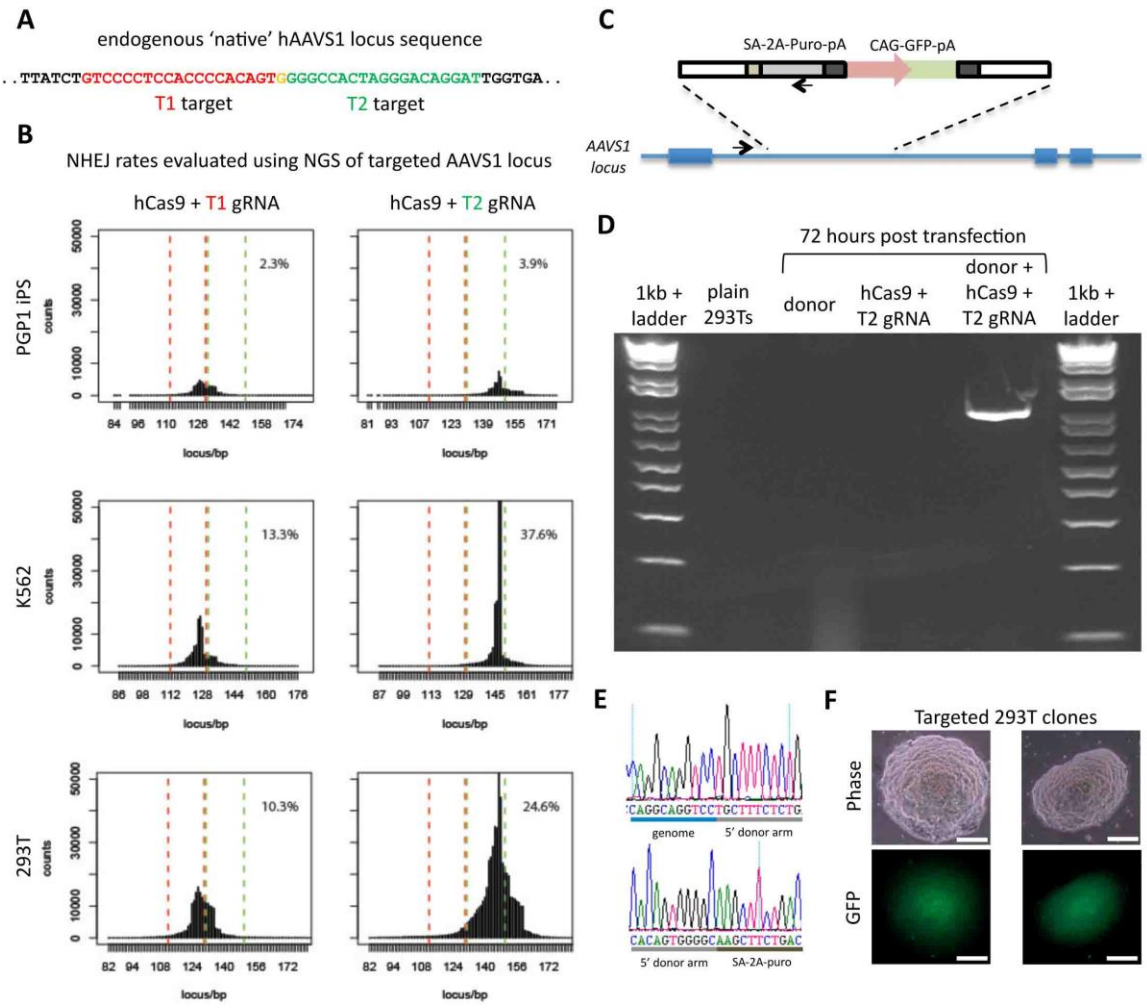


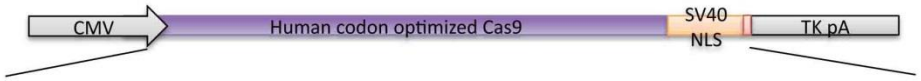
Figure 3-1. Genome editing in human cells using an engineered type II CRISPR system.

(A) RNA-guided gene targeting in human cells involves co-expression of the Cas9 protein bearing a C-terminus SV40 nuclear localization signal with one or more defined-length guide RNAs expressed from the human U6 polymerase III promoter. Cas9 unwinds the DNA duplex and cleaves both strands upon recognition of a target sequence by the guide RNA, but only if the correct protospacer-associated motif (PAM) is present at the 3' end. Any genomic sequence of the form GN20GG can in principle be targeted.

(B) A genomically integrated GFP coding sequence is disrupted by the insertion of a stop codon and a 68bp genomic fragment derived from the AAVS1 locus. Restoration of the GFP sequence by homologous recombination with an appropriate donor sequence results in GFP+ cells that can be quantitated by FACS. T1 and T2 gRNAs target sequences within the AAVS1 fragment. Binding sites for the two halves of the TAL effector nuclease heterodimer (TALEN) are underlined.

(C) Bar graph depicting homologous recombination efficiencies induced by T1, T2, and TALEN-mediated nuclease activity at the target locus, as measured by FACS. Representative FACS plots and microscopy images of the targeted cells are depicted below (scale bar is 100 microns).

A

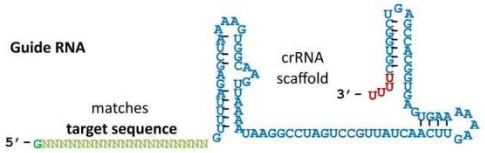


gccaccATGGACAAGAAGTACTCCATTGGGCTCGATATCGGCACAAACAGCGTCGGCTGGGCCGTCATTACGGACGAGTACAAGGTGCCGAGCAAAAAATTCAAAGTTCTGGGCAATACCGA
TCGCCACAGCATAAAGAAGAACCTCATTGGCGCCCTCCTGTTCGACTCCGGGGAGACGGCCGAAGCCACGCGGCTCAAAAGAACAGCACGGCGCAGATATACCCGCAGAAAGAATCGGATCT
GCTACCTGCAGGAGATCTTTAGTAATGAGATGGCTAAGGTGGATGACTCTTTCTTCCATAGGCTGGAGGAGTCCTTTTTGGTGGAGGAGGATAAAAAGCACGAGCGCCACCCAATCTTTGGC
AATATCGTGGACGAGGTGGCGTACCATGAAAAGTACCCAACCATATATCATCTGAGGAAGAAGCTTGTAGACAGTACTGATAAGGCTGACTTGCGGTTGATCTATCTCGCGCTGGCGCATAT
GATCAAATTTCGGGGACACTTCCTCATCGAGGGGGACCTGAACCCAGACAACAGCGATGTCGACAAACTCTTTATCCAACTGGTTCAGACTTACAATCAGCTTTTCGAAGAGAACCCGATCA
ACGCATCCGGAGTTGACGCCAAAGCAATCCTGAGCGCTAGGCTGTCCAAATCCCGGCGGCTCGAAAACCTCATCGCACAGCTCCCTGGGGAGAAGAAGAACGGCCTGTTTGGTAATCTTATC
GCCCTGTCACTCGGGCTGACCCCCAACTTTAAATCTAACTTCGACCTGGCCGAAGATGCCAAGCTTCAACTGAGCAAAGACACCTACGATGATGATCTCGACAATCTGCTGGCCCAGATCGG
CGACCAGTACGCAGACCTTTTTTTGGCGGCAAAGAACCTGTCAGACGCCATTCTGCTGAGTGATATTCTGCGAGTGAACACGGAGATCACCAAAGCTCCGCTGAGCGCTAGTATGATCAAGC
GCTATGATGAGCACCACCAAGACTTGACTTTGCTGAAGGCCCTTGTCAGACAGCAACTGCCTGAGAAGTACAAGGAAATTTTCTTCGATCAGTCTAAAAATGGCTACGCCGGATACATTGAC
GGCGGAGCAAGCCAGGAGGAATTTTACAAATTTATTAAGCCCATCTTGGAAAAAATGGACGGCACCGAGGAGCTGCTGGTAAAGCTTAACAGAGAAGATCTGTTGCGCAAACAGCGCACTTT
CGACAATGGAAGCATCCCCCACCAGATTCACCTGGGCGAACTGCACGCTATCCTCAGGCGGCAAGAGGATTTCTACCCCTTTTGAAAGATAACAGGGAAAAGATTGAGAAAATCCTCACAT
TTCGGATACCCTACTATGTAGGCCCCCTGCCCGGGGAAATTCCAGATTCGCGTGGATGACTCGCAAATCAGAAGAGACCATCACTCCCTGGAACTTCGAGGAAGTCGTGGATAAGGGGGCC
TCTGCCCAGTCCTTCATCGAAAGGATGACTAACTTTGATAAAAATCTGCCTAACGAAAAGGTGCTTCCTAAACACTCTCTGCTGTACGAGTACTTCACAGTTTATAACGAGCTCACCAAGGT
CAAATACGTCACAGAAGGGATGAGAAAGCCAGCATTCCTGTCTGGAGAGCAGAAGAAAGCTATCGTGGACCTCCTCTTCAAGACGAACCGGAAAGTTACCGTGAAACAGCTCAAAGAAGACT
ATTTCAAAAAGATTGAATGTTTCGACTCTGTTGAAATCAGCGGAGTGGAGGATCGCTTCAACGCATCCCTGGGAACGTATCACGATCTCCTGAAAATCATTAAAGACAAGGACTTCCTGGAC
AATGAGGAGAACGAGGACATTCTTGAGGACATTGTCCTCACCCTTACGTTGTTTGAAGATAGGGAGATGATTGAAGAACGCTTGAAAACTTACGCTCATCTCTTCGACGACAAAGTCATGAA
ACAGCTCAAGAGGCGCCGATATACAGGATGGGGGCGGCTGTCAAGAAAACTGATCAATGGGATCCGAGACAAGCAGAGTGGAAAGACAATCCTGGATTTTCTTAAGTCCGATGGATTTGCCA
ACCGGAACTTCATGCAGTTGATCCATGATGACTCTCTCACCTTTAAGGAGGACATCCAGAAAGCACAAGTTTCTGGCCAGGGGGACAGTCTTCACGAGCACATCGCTAATCTTGCAGGTAGC
CCAGCTATCAAAAAGGGAATACTGCAGACCGTTAAGGTCGTGGATGAACTCGTCAAAGTAATGGGAAGGCATAAGCCCGAGAATATCGTTATCGAGATGGCCCGAGAGAACCAAACTACCCA
GAAGGGACAGAAGAACAGTAGGGAAAGGATGAAGAGGATTGAAGAGGGTATAAAAGAACTGGGGTCCCAAATCCTTAAGGAACACCCAGTTGAAAACACCCAGCTTCAGAATGAGAAGCTCT
ACCTGTACTACCTGCAGAACGGCAGGGACATGTACGTGGATCAGGAACTGGACATCAATCGGCTCTCCGACTACGACGTGGATCATATCGTGCCCCAGTCTTTTCTCAAAGATGATTCTATT
GATAATAAAGTGTTGACAAGATCCGATAAAAATAGAGGGAAGAGTGATAACGTCCCCTCAGAAGAAGTTGTCAAGAAAATGAAAAATTATTGGCGGCAGCTGCTGAACGCCAAACTGATCAC
ACAACGGAAGTTCGATAATCTGACTAAGGCTGAACGAGGTGGCCTGTCTGAGTTGGATAAAGCCGGCTTCATCAAAAGGCAGCTTGTTGAGACACGCCAGATCACCAAGCACGTGGCCCAAA
TTCTCGATTCACGCATGAACACCAAGTACGATGAAAATGACAAACTGATTCGAGAGGTGAAAGTTATTACTCTGAAGTCTAAGCTGGTCTCAGATTTCAGAAAGGACTTTCAGTTTTATAAG
GTGAGAGAGATCAACAATTACCACCATGCGCATGATGCCTACCTGAATGCAGTGGTAGGCACTGCACTTATCAAAAATATCCCAAGCTTGAATCTGAATTTGTTTACGGAGACTATAAAGT
GTACGATGTTAGGAAAATGATCGCAAAGTCTGAGCAGGAAATAGGCAAGGCCACCGCTAAGTACTTCTTTTACAGCAATATTATGAATTTTTCAAGACCGAGATTACACTGGCCAATGGAG
AGATTCGGAAGCGACCACTTATCGAAACAAACGGAGAAACAGGAGAAATCGTGTGGGACAAGGGTAGGGATTTCGCGACAGTCCGGAAGGTCCTGTCCATGCCGCAGGTGAACATCGTTAAA
AAGACCGAAGTACAGACCGGAGGCTTCTCCAAGGAAAGTATCCTCCCGAAAAGGAACAGCGACAAGCTGATCGCACGCAAAAAAGATTGGGACCCCAAGAAATACGGCGGATTCGATTCTCC
TACAGTCGCTTACAGTGTACTGGTTGTGGCCAAAGTGGAGAAAGGGAAGTCTAAAAAACTCAAAAGCGTCAAGGAACTGCTGGGCATCACAATCATGGAGCGATCAAGCTTCGAAAAAACC
CCATCGACTTTCTCGAGGCGAAAGGATATAAAGAGGTCAAAAAAGACCTCATCATTAAGCTTCCCAAGTACTCTCTCTTTGAGCTTGAAAACGGCCGGAAACGAATGCTCGCTAGTGCGGGC
GAGCTGCAGAAAGGTAACGAGCTGGCACTGCCCTCTAAATACGTTAATTTCTTGTATCTGGCCAGCCACTATGAAAAGCTCAAAGGGTCTCCCGAAGATAATGAGCAGAAGCAGCTGTTCGT
GGAACAACACAAACACTACCTTGATGAGATCATCGAGCAAATAAGCGAATTCTCCAAAAGAGTGATCCTCGCCGACGCTAACCTCGATAAGGTGCTTTCTGCTTACAATAAGCACAGGGATA
AGCCCATCAGGGAGCAGGCAGAAAACATTATCCACTTGTTTACTCTGACCAACTTGGGCGCGCCTGCAGCCTTCAAGTACTTCGACACCACCATAGACAGAAAGCGGTACACCTCTACAAAG
GAGGTCCTGGACGCCACACTGATTCATCAGTCAATTACGGGGCTCTATGAAACAAGAATCGACCTCTCTCAGCTCGGTGGAGACAGCAGGGCTGACCCCAAGAAGAAGAGGAAGGTGTGA

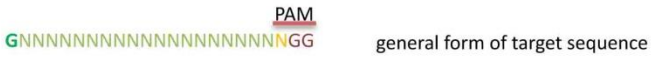
B

U6 promoter + target RNA + guide RNA scaffold:

TGTACAAAAAAGCAGGCTTTAAAGGAACCAATTCAGTCGACTGGATCCGGTACCAAGGTCGGGCAGGAAGAGGGCCTATTTCCCATGAT
TCCTTCATATTTGCATATACGATACAAGGCTGTTAGAGAGATAATTAGAATTAATTTGACTGTAAACACAAAGATATTAGTACAAAATA
CGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAAATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGA
AAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGNNNNNNNNNNNNNNNNNNNNGTTTTAGAGCTAGAAATA
GCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTTCTAGACCCAGCTTTCTTGTACAAA
GTTGGCATTA



C



Name	Target Sequence
GFP gRNA Target 1	GTGAACCGCATCGAGCTGAAGGG
GFP gRNA Target 2	GGAGCGCACCATCTTCTTCAAGG
AAVS1 gRNA Target 1	GTCCCCTCCACCCCACAGTGGGG
AAVS1 gRNA Target 2	GGGGCCACTAGGGACAGGATTGG

Figure 3_2. The engineered type II CRISPR system for human cells. (continued)

(A) Expression format and full sequence of the cas9 gene insert. The RuvC-like and HNH motifs, and the C-terminus SV40 NLS are respectively highlighted by blue, brown and orange colors.

(B) U6 promoter based expression scheme for the guide RNAs and predicted RNA transcript secondary structure. The use of the U6 promoter constrains the 1st position in the RNA transcript to be a 'G' and thus all genomic sites of the form GN20GG can be targeted using this approach.

(C) A list of the 4 gRNAs used in this study for targeting GFP and the AAVS1 locus is provided**.**

To test the functionality of our implementation for genome engineering, we developed a green fluorescent protein (GFP) reporter assay (Figure 3_1 B) in human embryonic kidney HEK 293T cells similar to one previously described (*13*). Specifically, we established a stable cell line bearing a genomically integrated GFP coding sequence disrupted by the insertion of a stop codon and a 68-bp genomic fragment fromtheAAVS1 locus that renders the expressed protein fragment non- fluorescent. Homologous recombination (HR) using an appropriate repair donor can restore the normal GFP sequence, which enabled us to quantify the resulting GFP+ cells by flow-activated cell sorting (FACS).

To test the efficiency of our system at stimulating HR, we constructed two gRNAs, T1 and T2, that target the intervening AAVS1 fragment (Figure 3_2 B) and compared their activity to that of a previously described TAL effector nuclease heterodimer (TALEN) targeting the same region (*14*).We observed successful HR events using all three targeting reagents, with gene correction rates using the T1 andT2 gRNAs approaching 3%and 8%, respectively (Figure 3_2 C). This RNA-mediated editing process was notably rapid, with the first detectable GFP+ cells appearing ~20 hours post transfection compared with ~40 hours for the AAVS1 TALENs. We observed HR only upon simultaneous introduction of the repair donor, Cas9 protein, and gRNA, which confirmed that all components are required for genome editing (Figure 3_3). Although we noted no apparent toxicity associated with Cas9/gRNA expression, work with zinc finger nucleases (ZFNs) and TALENs has shown that nicking only one strand further reduces toxicity. Accordingly, we also tested a Cas9D10A mutant that is known to function as a nickase in vitro, which yielded similar HR but lower non homologous end

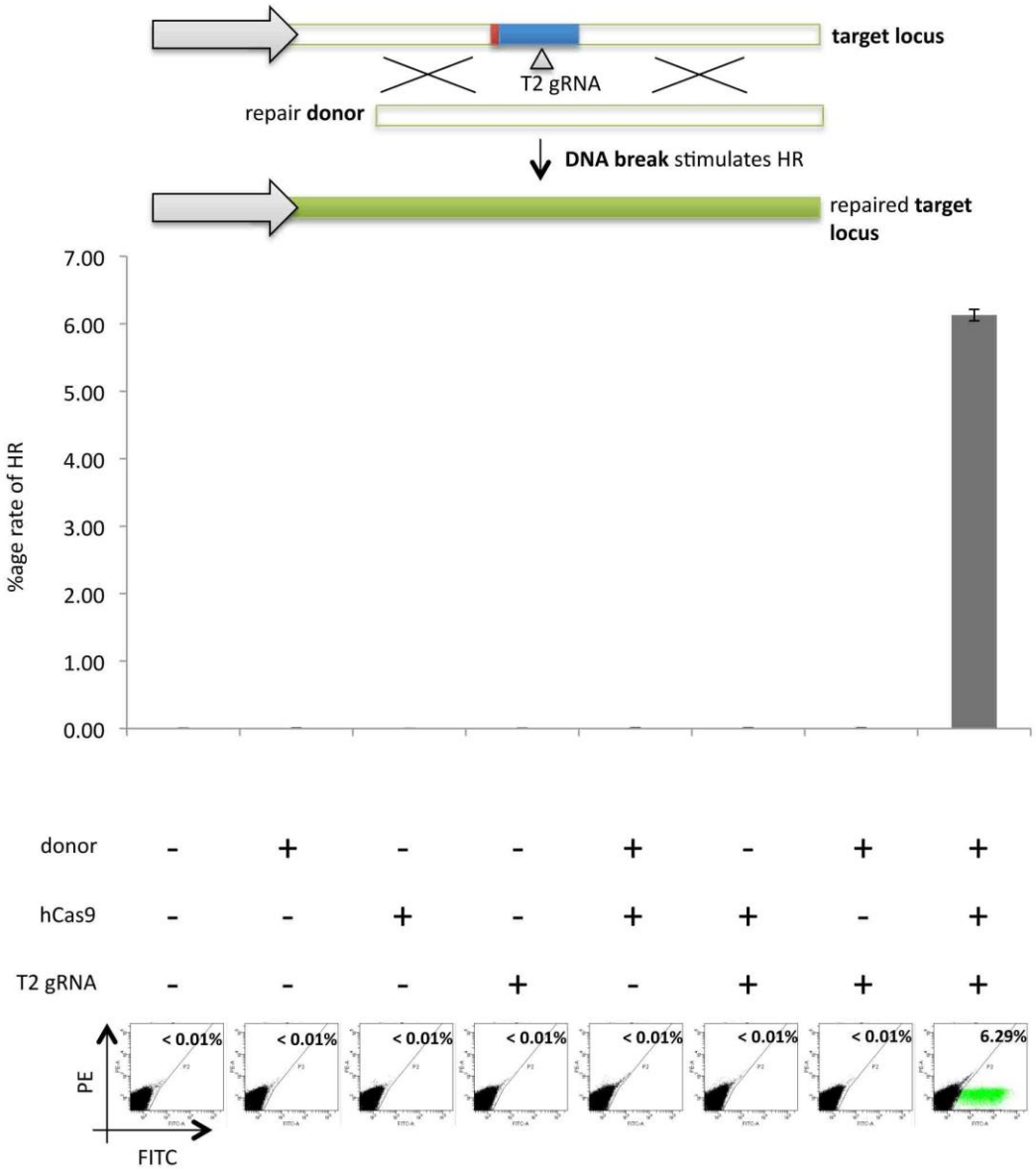


Figure 3_3. RNA-guided genome editing requires both Cas9 and guide RNA for successful targeting. Using the GFP reporter assay described in Fig. 1B, all possible combinations of the repair DNA donor, Cas9 protein, and gRNA were tested for their ability to effect successful HR. GFP+ cells were observed only when all the 3 components were present, validating that these CRISPR components are essential for RNA-guided genome editing.

joining (NHEJ) rates (Figure 3_4) (*2*), Consistent with (*2*), in which a related Cas9 protein is shown to cut both strands 3 bp upstream of the PAM, our NHEJ data confirmed that most deletions or insertions occurred at the 3′ end of the target sequence (Figure 3_4 B). We also confirmed that mutating the target genomic site prevents the gRNA from effecting HR at that locus, which demonstrates that CRISPR- mediated genome editing is sequence-specific (Figure 3_5). Finally, we showed that two gRNAs targeting sites in the GFP gene, and also three additional gRNAs targeting fragments from homologous regions of the DNA methyl transferase 3a (DNMT3a) and DNMT3b genes could sequence-specifically induce significant HR in the engineered reporter cell lines (Figure 3_6 and Figure 3_7). Together, these results confirm that RNA- guided genome targeting in human cells is simple to execute and induces robustHRacrossmultiple target sites.

Having successfully targeted an integrated reporter, we next turned to modifying a native locus. We used the gRNAs described above to target theAAVS1 locus located in the PPP1R12C gene on chromosome 19, which is ubiquitously expressed across most tissues (Figure 3_8 A).We targeted 293Ts, human chronic myelogenous leukemia K562 cells, and PGP1 human induced pluripotent stem (iPS) cells (*15*) and analyzed the results by next-generation sequencing of the targeted locus. Consistent with our results for the GFP reporter assay, we observed high numbers of NHEJ events at the endogenous locus for all three cell types. The two gRNAs T1 and T2 achieved NHEJ rates of 10 and 25%in 293Ts, 13 and 38% in K562s, and 2 and 4% in PGP1- iPS cells, respectively (Figure 3_8). We observed no overt toxicity from the Cas9 and gRNA expression required to induce NHEJ in any of these cell types. As expected, NHEJ-mediated deletions for T1 and T2 were centered around the target site positions, which further validated the

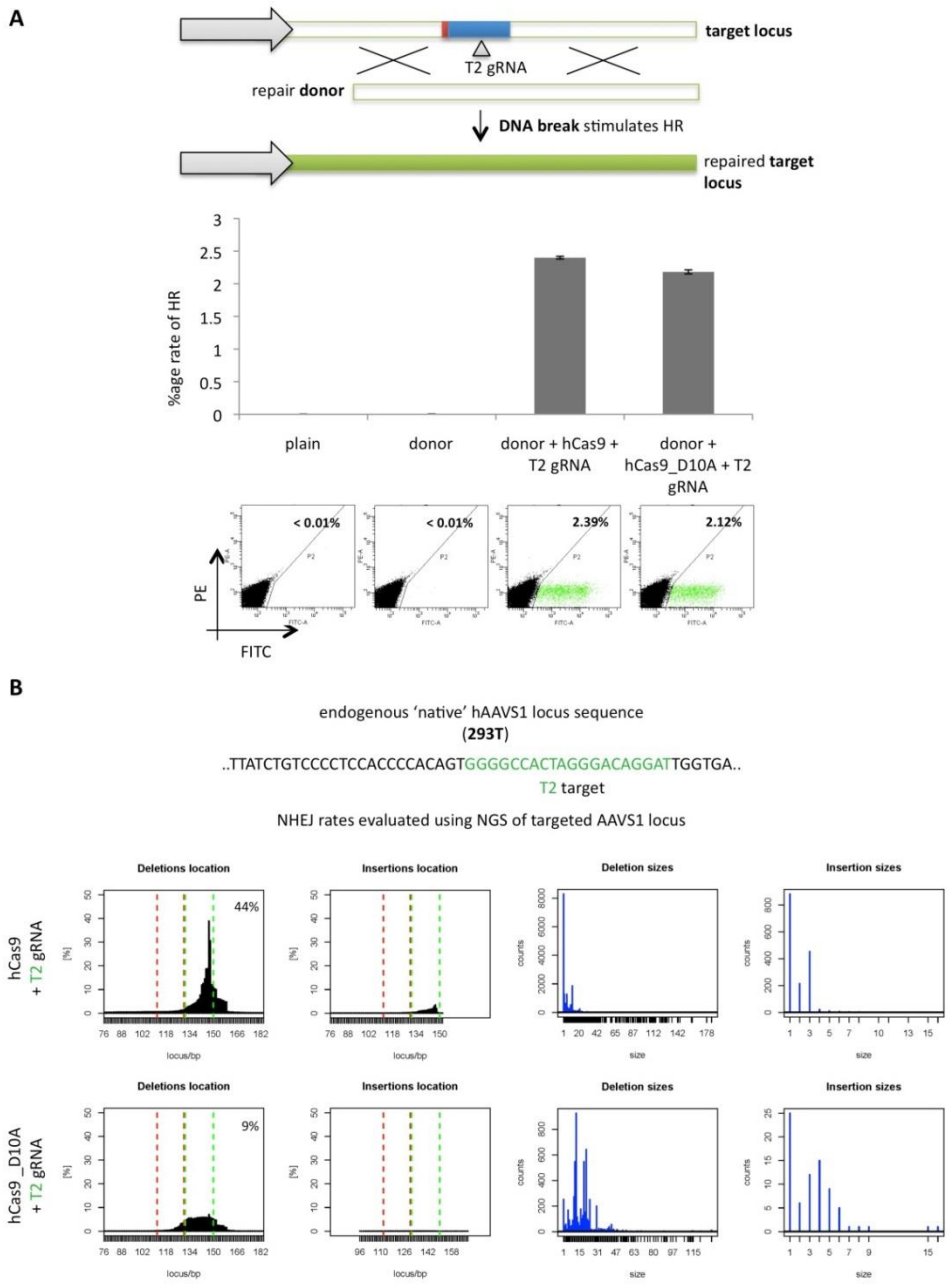


Figure 3_4. Analysis of gRNA and Cas9 mediated genome editing. We closely examined the CRISPR mediated genome editing process using either
(A) a GFP reporter assay as described earlier, and
(B) deep sequencing of the targeted loci. As comparison we also tested a D10A mutant for Cas9 that has been shown in earlier reports to function as a nickase in in vitro assays. Our data shows that both Cas9 and Cas9D10A can effect successful HR at nearly similar rates. Deep sequencing however confirms that while Cas9 shows robust NHEJ at the targeted loci, the D10A mutant has significantly diminished NHEJ rates (as would be expected from its putative ability to only nick DNA). Also, consistent with the known biochemistry of the Cas9 protein, our NHEJ data confirms that most base-pair deletions or insertions occurred near the 3' end of the target sequence: the peak is ~3-4 bases upstream of the PAM site, with a median deletion frequency of ~9-10bp. Data is mean +/- SEM (N=3).

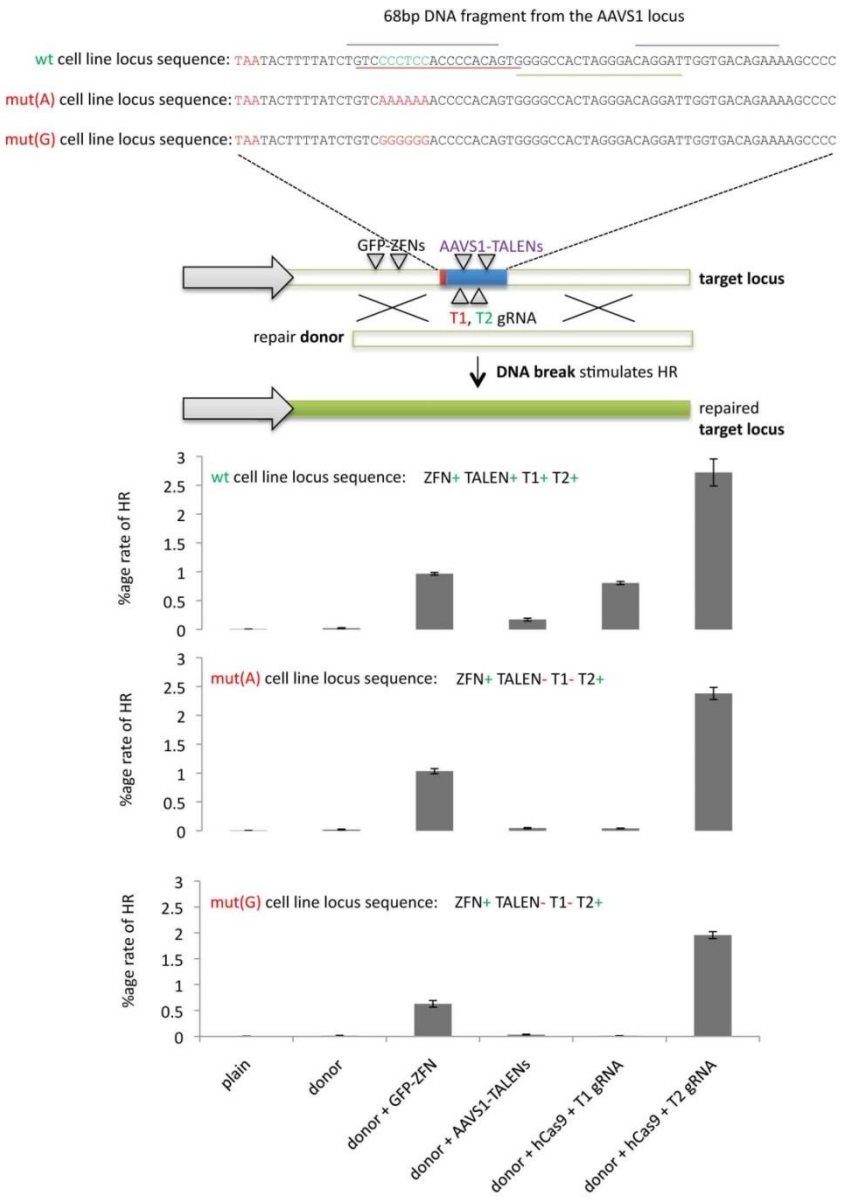


Figure 3_5. RNA-guided genome editing is target sequence specific.
we developed 3 293T stable lines each bearing a distinct GFP reporter construct. These are distinguished by the sequence of the AAVS1 fragment insert (as indicated in the figure). One line harbored the wild-type fragment while the two other lines were 6 mutated bases away (highlighted in red). Each of the lines was then targeted by one of the following 4 reagents: a GFP-ZFN pair that can target all cell types since its targeted sequence was in the flanking GFP fragments and hence present in along cell lines; a AAVS1 TALEN that could potentially target only the wt-AAVS1 fragment since the mutations in the other two lines should render the left TALEN unable to bind their sites; the T1 gRNA which can also potentially target only the wt-AAVS1 fragment, since its target site is also disrupted in the two mutant lines; and finally the T2 gRNA which should be able to target all the 3 cell lines since unlike the T1 gRNA its target site is unaltered among the 3 lines. Consistent with these predictions, the ZFN modified all 3 cell types, the AAVS1 TALENs and the T1 gRNA only targeted the wt-AAVS1 cell type, and the T2 gRNA successfully targets all 3 cell types. These results together confirm that the guide RNA mediated editing is target sequence specific.

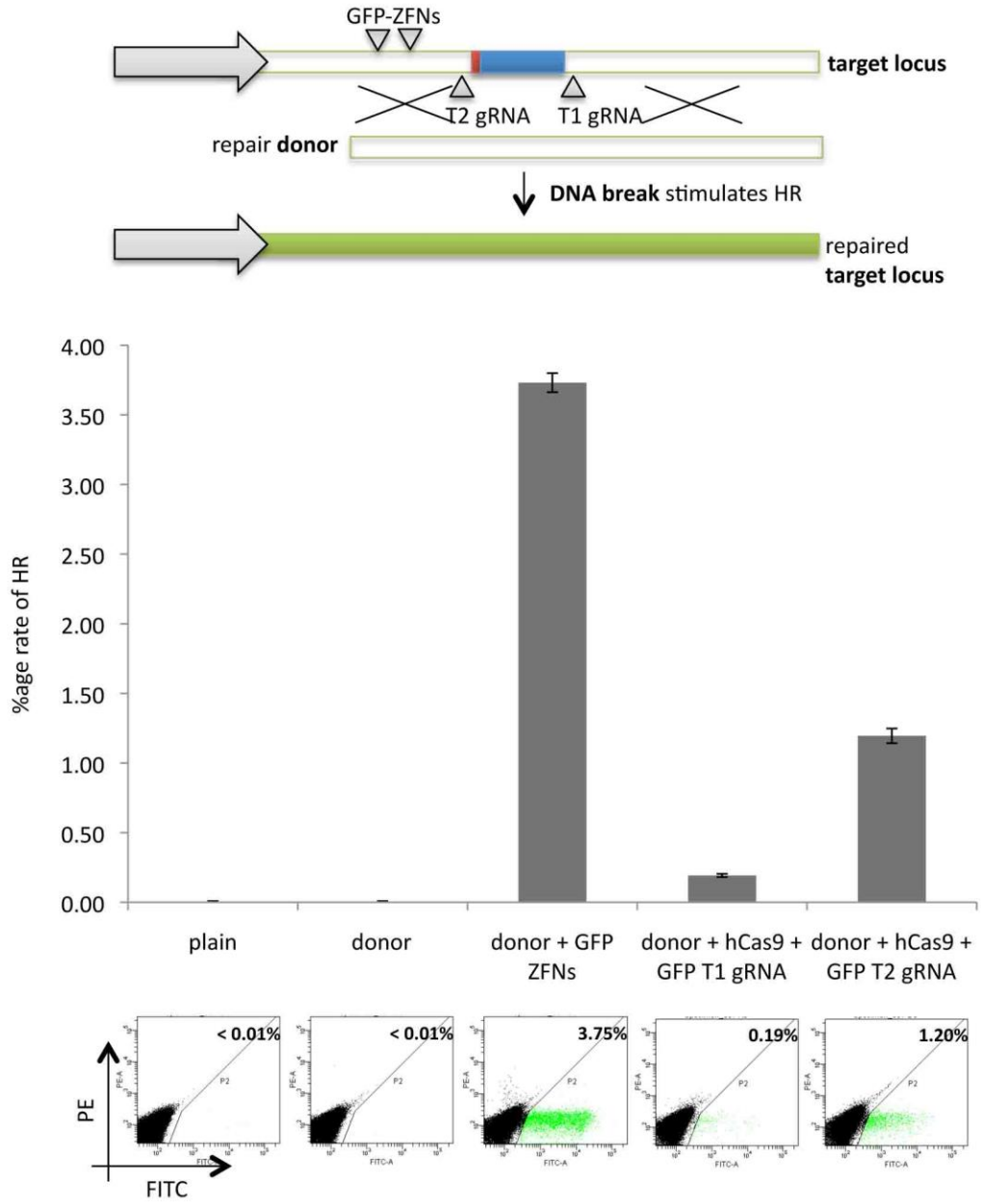


Figure 3_6. Guide RNAs targeted to the GFP sequence enable robust genome editing.
In addition to the 2 gRNAs targeting the AAVS1 insert, we also tested two additional gRNAs targeting the flanking GFP sequences of the reporter described in Fig. 1B. These gRNAs were also able to effect robust HR at this engineered locus.

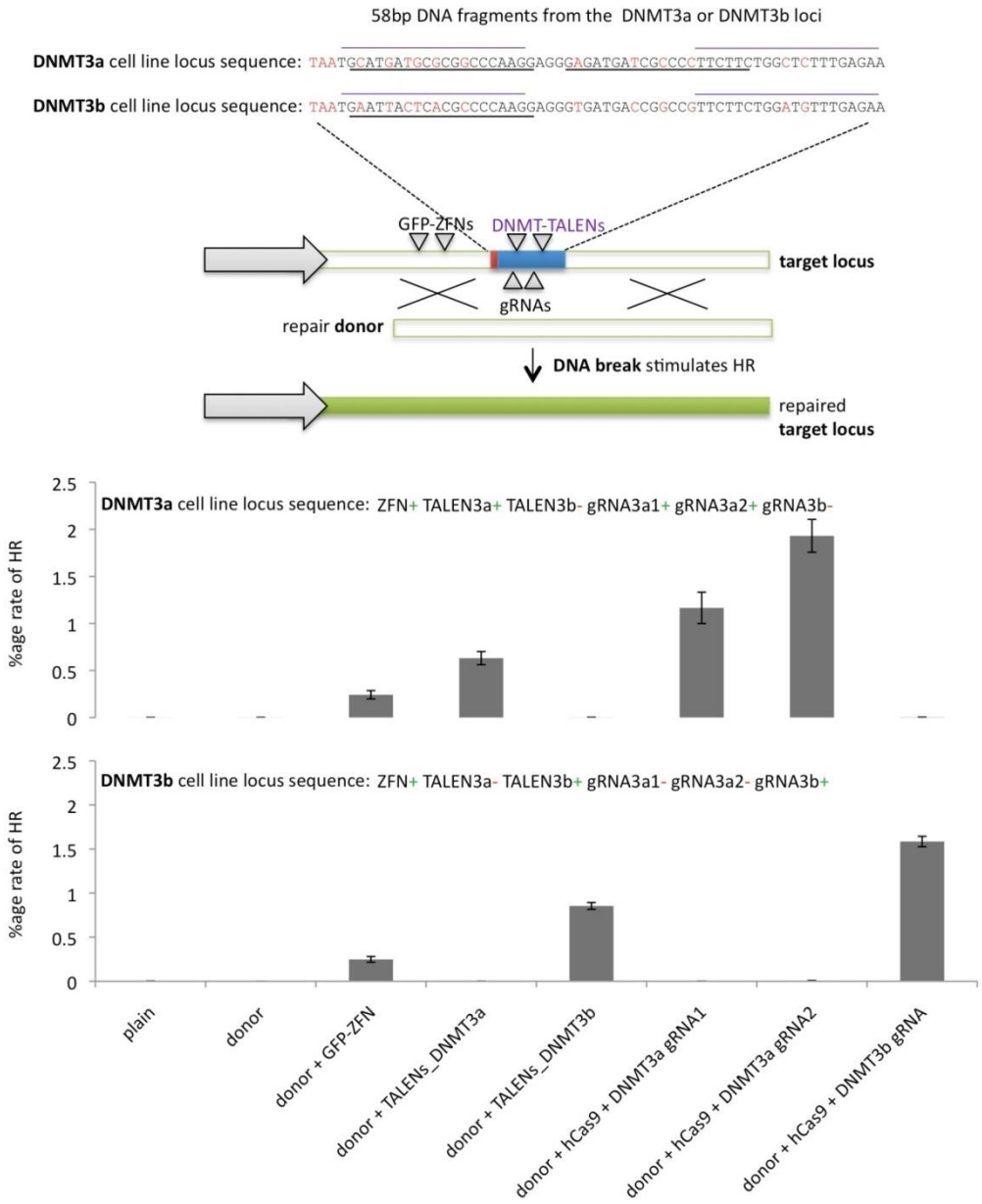


Figure 3_7. RNA-guided genome editing is target sequence specific, and demonstrates similar targeting efficiencies as ZFNs or TALENs.

Similar to the GFP reporter assay described in Fig. 1B, we developed 2 293T stable lines each bearing a distinct GFP reporter construct. These are distinguished by the sequence of the fragment insert (as indicated in the figure). One line harbored a 58bp fragment from the DNMT3a gene while the other line bore a homologous 58bp fragment from the DNMT3b gene. The sequence differences are highlighted in red. Each of the lines was then targeted by one of the following 6 reagents: a GFP-ZFN pair that can target all cell types since its targeted sequence was in the flanking GFP fragments and hence present in along cell lines; a pair of TALENs that potentially target either DNMT3a or DNMT3b fragments; a pair of gRNAs that can potentially target only the DNMT3a fragment; and finally a gRNA that should potentially only target the DNMT3b fragment. Consistent with these predictions, the ZFN modified all 3 cell types, and the TALENs and gRNAs only their respective targets. Furthermore the efficiencies of targeting were comparable across the 6 targeting reagents. These results together confirm that RNA-guided editing is target sequence specific and demonstrates similar targeting efficiencies as ZFNs or TALENs. Data is mean +/- SEM (N=3).

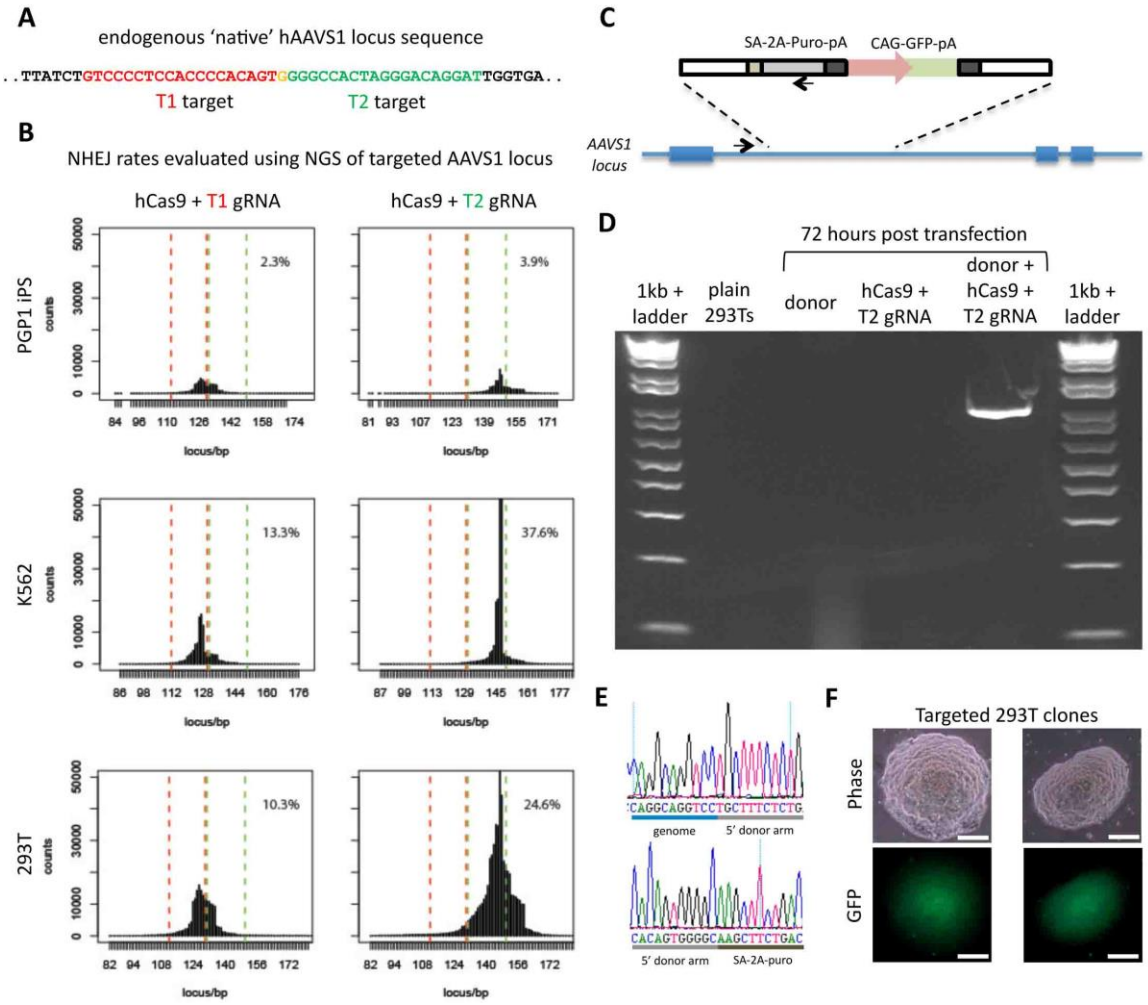


Figure 3_8. RNA-guided genome editing of the native AAVS1 locus in multiple cell types.

(A) Sequences targeted by T1 (red) and T2 (green) gRNAs are located inside an intron of PPP1R12C gene within the AAVS1 locus on chromosome 19.

(B) T1 and T2 gRNAs induced Cas9 to cleave target sequences within 293Ts, K562s, and PGP1 human iPS cells, resulting in NHEJ-mediated deletions that were pinpointed and quantified using next-generation sequencing. NHEJ frequencies for T1 and T2 gRNAs were 10% and 25% in 293T, 13% and 38% in K562, and 2% and 4% in PGP1 iPS cells, respectively. Red dash lines demarcate the boundary of the T1 gRNA targeting site; green dash lines demarcate the boundary of the T2 gRNA targeting site. Deletion incidences at each nucleotide position are plotted in black lines. The sequence of the whole targeting region which was used as the reference for NGS mapping is listed in supplement. As expected (refer Fig. 1A), the peak of frequency of NHEJ based base-pair deletions occurs at the 3 ' end of the target sequence.

(C) DNA donor architecture for HR at the AAVS1 locus, and location of the sequencing primers (arrows) to detect successful targeted events is depicted.

(D) PCR assay three days post transfection demonstrates that only cells expressing the donor, Cas9 and T2 gRNA show evidence of successful HR events.

(E) Successful HR was confirmed by Sanger sequencing of the PCR amplicon showing that the expected DNA bases at both the genome-donor and donor-insert boundary are present.

(F) Successfully targeted clones of 293T cells were also selected with puromycin for 2 weeks. Microscope images of two representative GFP+ clones is shown (scale bar is 100 microns).

sequence- specificity of this targeting process (Figure 3_9-11). Simultaneous introduction of both T1 and T2 gRNAs resulted in high-efficiency deletion of the intervening 19-bp fragment (Figure 3_10), which demonstrated that multiplexed editing of genomic loci is feasible using this approach. Last, we attempted to use HR to integrate either a double-stranded DNA donor construct (*16*) or an oligo donor into the native AAVS1 locus (Figure 3_8 and Figure 3_12).We confirmed HR- mediated integration, using both approaches, by polymerase chain reaction (PCR) (Figure 3_8 D and Figure 3_12) and Sanger sequencing (Figure 3_8 E). We also readily derived 293T or iPS clones from the pool ofmodified cells using puromycin selection over 2 weeks (Figure 3_8 E and Figure 3_12). These results demonstrate that this approach enables efficient integration of foreign DNA at endogenous loci in human cells.

Comparison of Cas9-gRNA system with reTALENs system in human stem cells

After confirming activity of Cas9-gRNA, we next sought to compare the efficiency reTALENs V.S. Cas9-gRNA . To do that, we design and constructed reTALENs and Cas9-gRNAs targeted to fifteen sites at the CCR5 genomic locus (Figure 3_13 A). Anticipating that editing efficiency might depend on chromatin state, these sites were selected to represent a wide range of DNaseI sensitivities (*17*). The nuclease constructs were transfected with the corresponding ssODNs donors (Supplementary Table 3) into PGP1 hiPSCs. Six days after transfection, we profiled the genome editing efficiencies at these sites (Supplementary Table 4). For 13 out of 15 re-TALEN pairs with ssODN donors, we detected NHEJ and HDR at levels above our statistical detection thresholds, with an average NHEJ efficiency of 0.4% and an average HDR efficiency of

Figure 3_9. RNA-guided NHEJ in human iPS cells.
we measured NHEJ rate by assessing genomic deletion and insertion rate at double-strand breaks (DSBs) by deep sequencing. *Panel 1*: Deletion rate detected at targeting region. Red dash lines: boundary of T1 RNA targeting site; green dash lines: boundary of T2 RNA targeting site. We plot the deletion incidence at each nucleotide position in black lines and we calculated the deletion rate as the percentage of reads carrying deletions. *Panel 2*: Insertion rate detected at targeting region. Red dash lines: boundary of T1 RNA targeting site; green dash lines: boundary of T2 RNA targeting site. We plot the incidence of insertion at the genomic location where the first insertion junction was detected in black lines and we calculated the insertion rate as the percentage of reads carrying insertions. *Panel 3*: Deletion size distribution. We plot the frequencies of different size deletions among the whole NHEJ population. *Panel 4*: insertion size distribution. We plot the frequencies of different sizes insertions among the whole NHEJ population.
(Continued)

endogenous 'native' hAAVS1 locus sequence
(**PGP1 iPS**)

..TTATCTGTCCCCTCCACCCCACAGTGGGGCCACTAGGGACAGGATTGGTGA..
T1 target T2 target

NHEJ rates evaluated using NGS of targeted AAVS1 locus

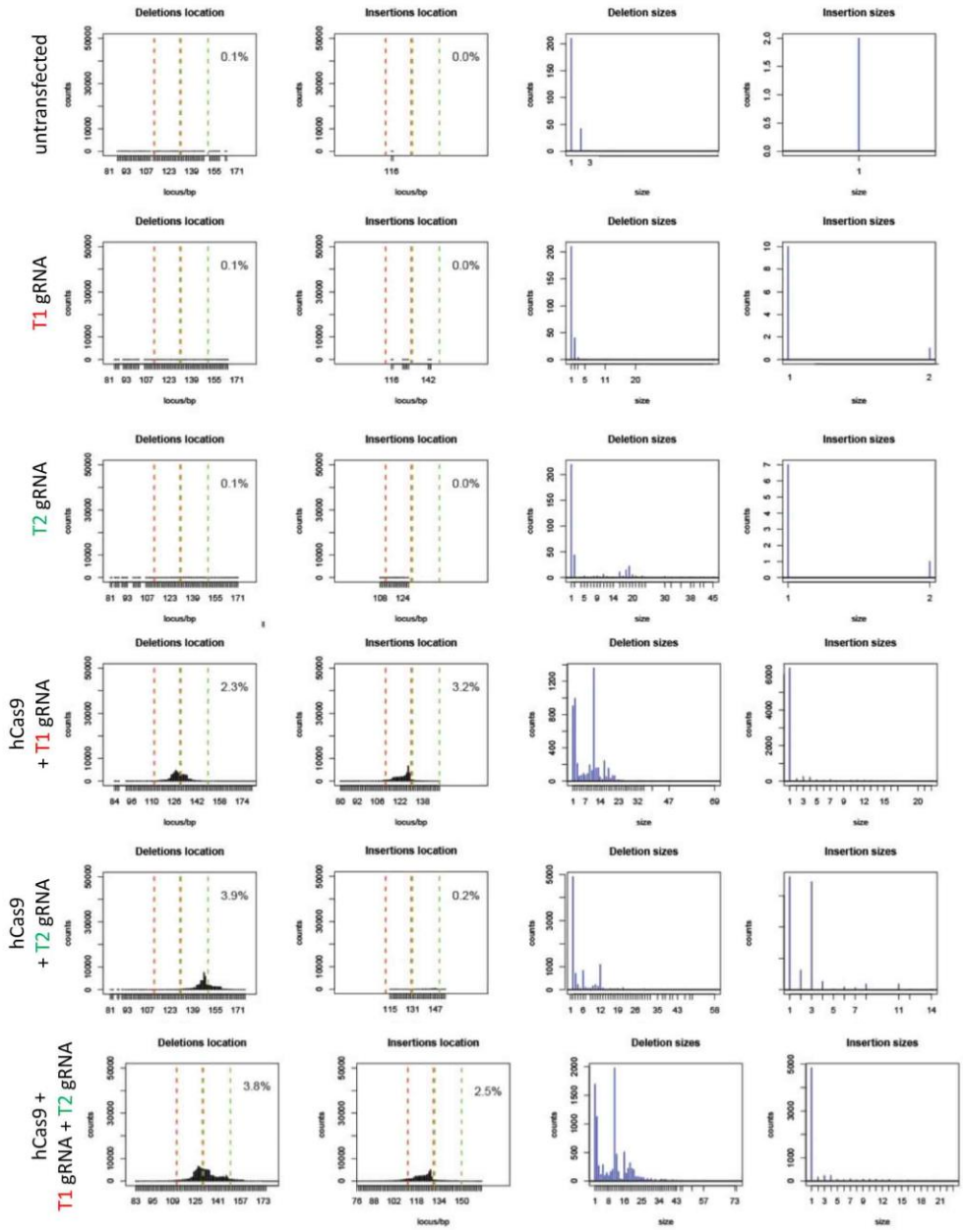


Figure 3_9 (Continued)

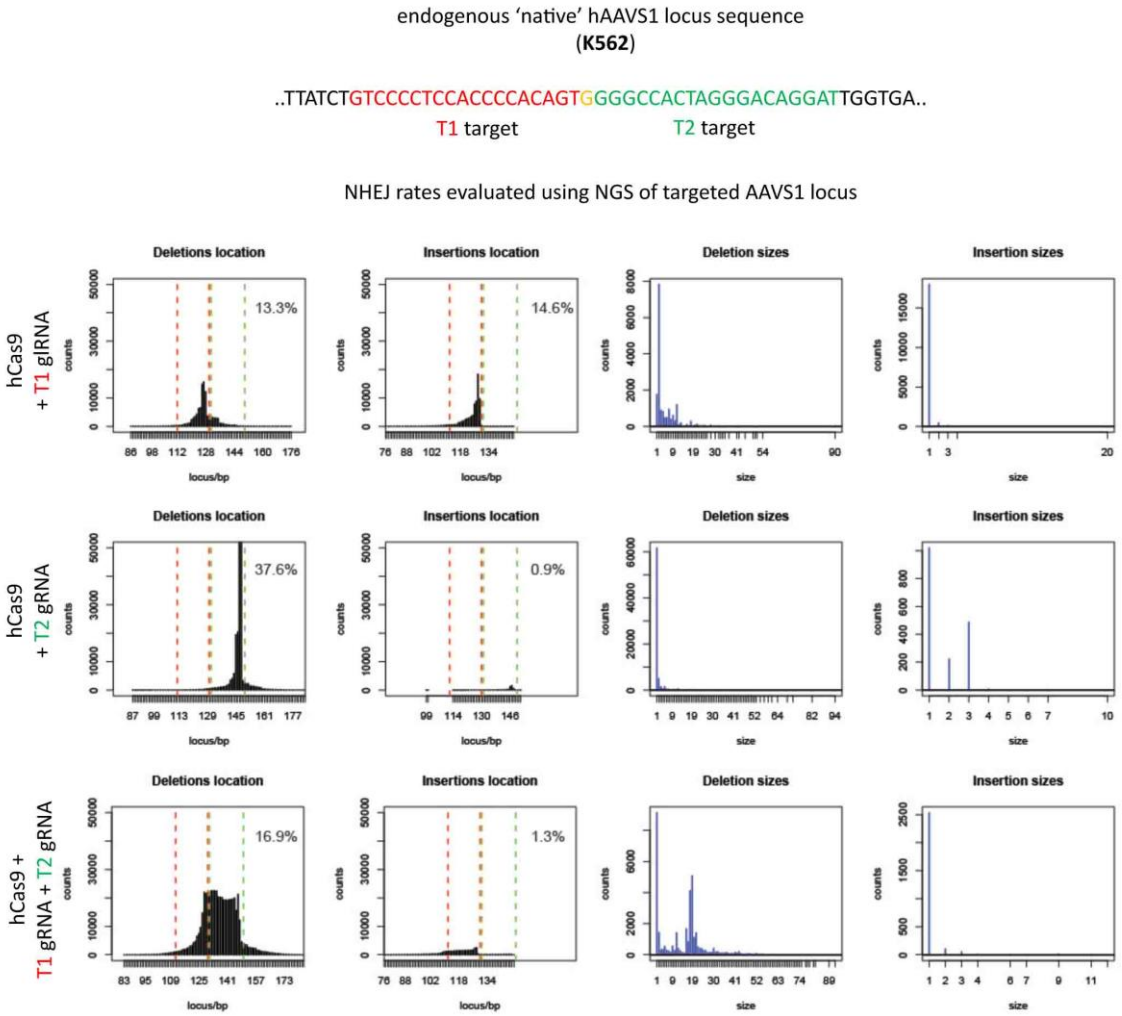


Figure 3_10. RNA-guided NHEJ in K562 cells.
K562 targeting by both gRNAs is efficient (13-38%) and sequence specific (as shown by the shift in position of the NHEJ deletion distributions). Importantly, as evidenced by the peaks in the histogram of observed frequencies of deletion sizes, simultaneous introduction of both T1 and T2 guide RNAs resulted in high efficiency deletion of the intervening 19bp fragment, demonstrating that multiplexed editing of genomic loci is also feasible using this approach.

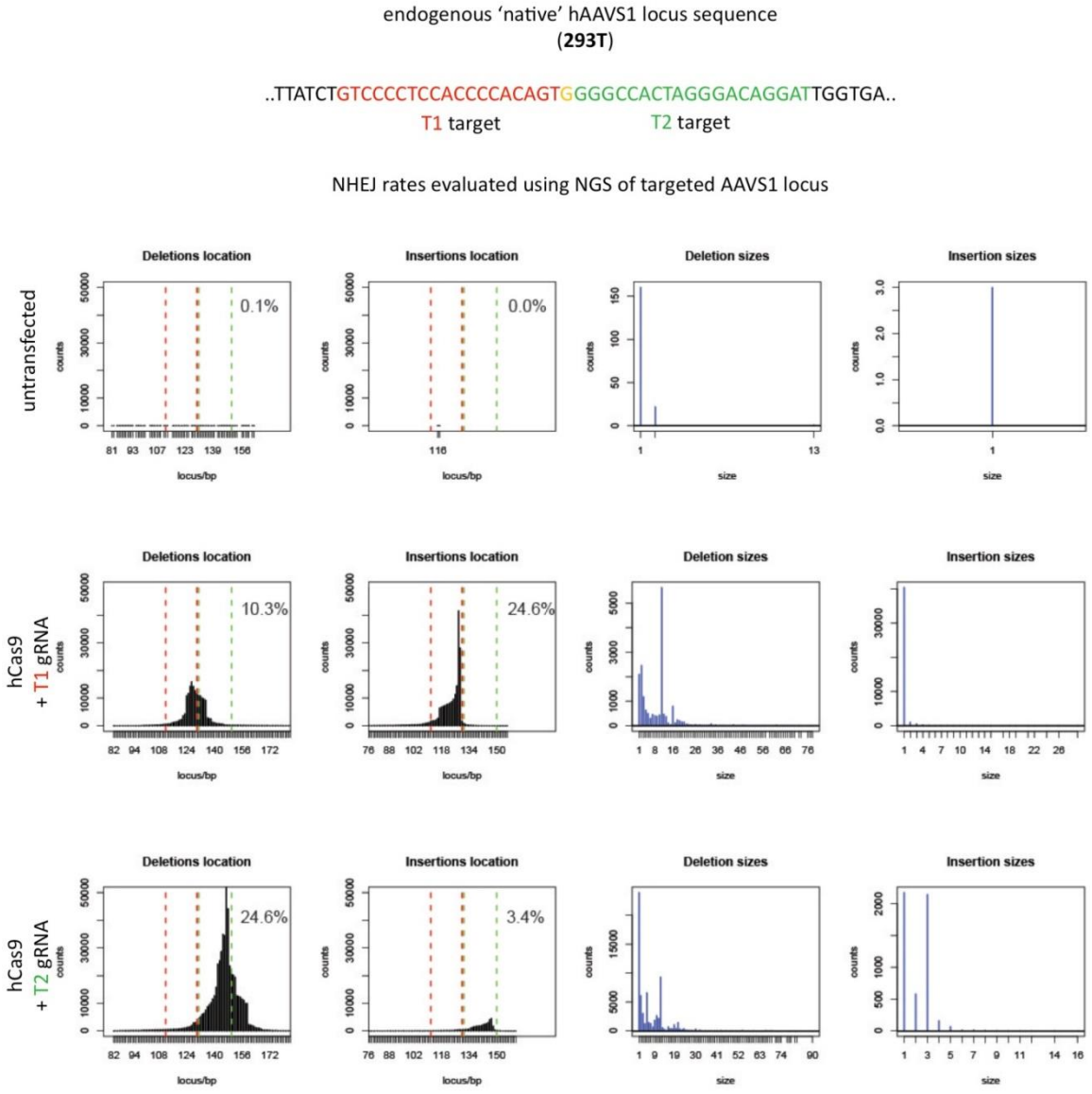


Figure 3_11. RNA-guided NHEJ in 293cells.
293T targeting by both gRNAs is efficient (10-24%) and sequence specific (as shown by the shift in position of the NHEJ deletion distributions).

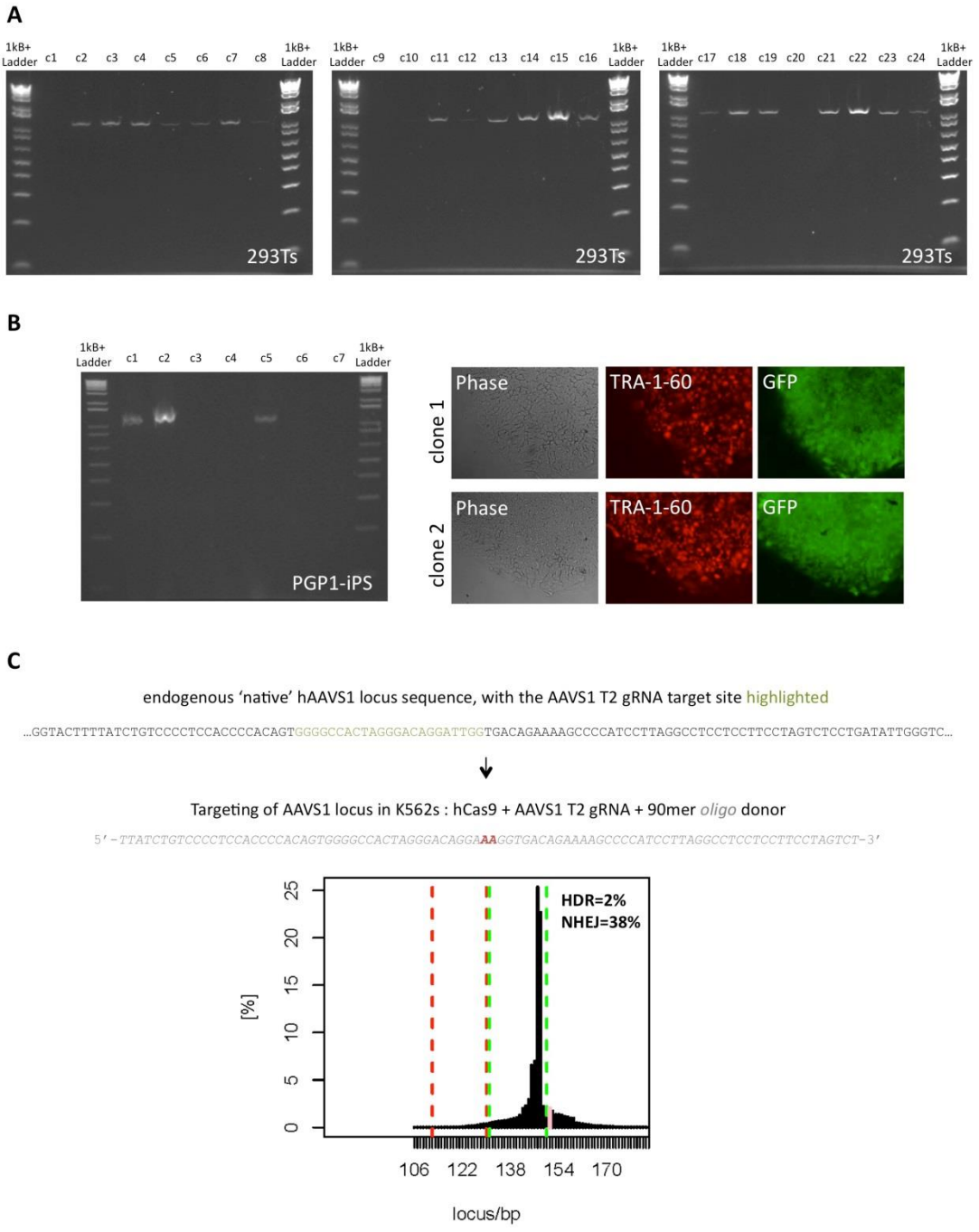


Figure 3_12. HR at the endogenous AAVS1 locus using either a dsDNA donor or a short oligonucleotide donor.
(A) PCR screen (refer Fig. 2C) confirmed that 21/24 randomly picked 293T clones were successfully targeted.
(B) Similar PCR screen confirmed 3/7 randomly picked PGP1-iPS clones were also successfully targeted.
(C) Finally short 90mer oligos could also effect robust targeting at the endogenous AAVS1 locus (shown here for K562 cells).

0.6% (Figure 3_13). In addition, a statistically significant positive correlation (r^2 =0.81) was found between HR and NHEJ efficiency at the same targeting loci ($P<1 \times 10^{-4}$) , suggesting that DSB generation, the common upstream step of both HDR and NHEJ, is a rate-limiting step for reTALEN-mediated genome editing.

In contrast, all 15 Cas9-gRNA pairs showed significant levels of NHEJ and HR, with an average NHEJ efficiency of 3% and an average HDR efficiency of 1.0% (Figure 3_13 A). In addition, a positive correlation was also detected between the NHEJ and HDR efficiency introduced by Cas9-gRNA (r^2=0.52, p=0.003), consistent with what we had observed with our reTALENs. The NHEJ efficiency achieved by Cas9-gRNA was significantly higher than that achieved by reTALENs (*t*-test, paired-end, P=0.02). Interestingly, we observed a moderate but statistically significant correlation between NHEJ efficiency and the melting temperature of the gRNA targeting sequence (Figure 3_13 B) (r^2=0.28, p=0.04), suggesting that the strength of base-pairing between the gRNA and its genomic target could explain as much as 28% of the variation in the efficiency of Cas9-gRNA-mediated DSB generation. Even though Cas9-gRNA produced NHEJ levels at an average of 7 times higher than the corresponding reTALEN, Cas9-gRNA only achieved HDR levels (average=1.0%) similar to that of the corresponding reTALENs (average = 0.6%), suggesting either that the ssODN concentration at the DSB is the limiting factor for HDR or that the genomic break structure created by the Cas9-gRNA is not favorable for effective HDR (See discussion). Of note, within our data, we did not observe any correlation between DNaseI HS and the genome targeting efficiencies achieved by either method.

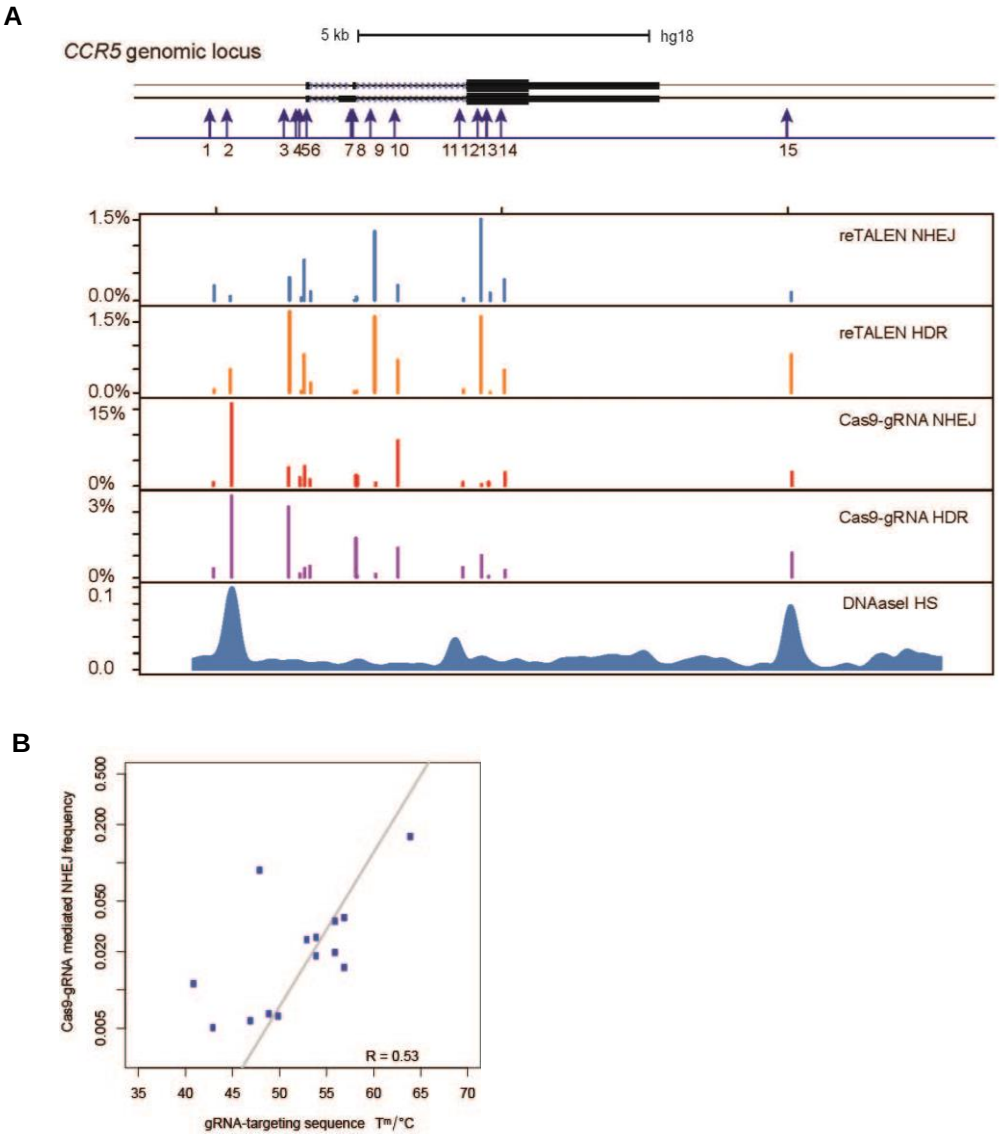


Figure 3_13. Comparison of the reTALEN and CRISPR activity

(A) The genome editing efficiency of re-TALENs and Cas9-gRNAs targeting *CCR5* in PGP1 hiPSCs.

Top: schematic representation of the targeted genome editing sites in *CCR5*. The 15 targeting sites are illustrated by blue arrows below. For each site, cells were co-transfected with a pair of re-TALENs and their corresponding ssODN donor carrying 2bp mismatches against the genomic DNA. Genome editing efficiencies were assayed 6 days after transfection. Similarly, we transfected 15 Cas9-gRNAs with their corresponding ssODNs individually into PGP1-hiPSCs to target the same 15 sites and analyzed the efficiency 6 days after transfection. Bottom: the genome editing efficiency of re-TALENs and Cas9-gRNAs targeting CCR5 in PGP1 hiPSCs. Panel 1 and 2 indicate NHEJ and HDR efficiencies mediated by reTALENs. Panel 3 and 4 indicate NHEJ and HDR efficiencies mediated by Cas9-gRNAs. NHEJ rates were calculated by the frequency of genomic alleles carrying deletions or insertions at the targeting region; HDR rates were calculated by the frequency of genomic alleles carrying 2bp mismatches. Panel 5, the DNaseI HS profile of a hiPSC cell line from ENCODE database (Duke DNase HS, iPS NIHi7 DS). Of note, the scales of different panels are different.

(B) The correlation of NHEJ efficiencies mediated by Cas9-gRNA and the Tm temperature of gRNA targeting site in iPSCs (r=0.52, P=0.04)

Computational design of gRNA array targeting the whole exome of human genome

Our versatile RNA-guided genome-editing system can be readily adapted to modify other genomic sites by simply modifying the sequence of our gRNA expression vector to match a compatible sequence in the locus of interest. To facilitate this process, we bioinformatically generated ~190,000 specific gRNA-targetable sequences targeting ~40.5% exons of genes in the human genome. We also incorporated these target sequences into a 200-bp format compatible with multiplex synthesis on DNA arrays (*18*) (Figure 3_14). This resource provides a ready genome-wide reference of potential target sites in the human genome and amethodology formultiplex gRNA synthesis.

Investigate the specificity of Cas9-gRNA and reTALEN

The ability to both edit and regulate genes using the above RNA-guided system opens the door to versatile multiplex genetic and epigenetic engineering of human cells. However, an increasingly recognized constraint on Cas9-mediated engineering is the apparently limited specificity of Cas9-gRNA targeting (*19*). Resolution of this issue will require in-depth interrogation of Cas9 affinity for a very large space of target sequence variations. We adapted our RNA-guided transcriptional activation system (hCRISPR-TF) we published recently to serve this purpose. This system provides a direct high-throughput readout of Cas9 targeting in human cells, avoids complications introduced by dsDNA cut toxicity and mutagenic repair incurred by specificity testing with native nuclease-active Cas9, and additionally can be adapted to any programmable DNA binding system. To illustrate this latter point, we also applied this system to evaluate TALE specificity. The methodology of our approach is outlined in Figure 3_15. Briefly,

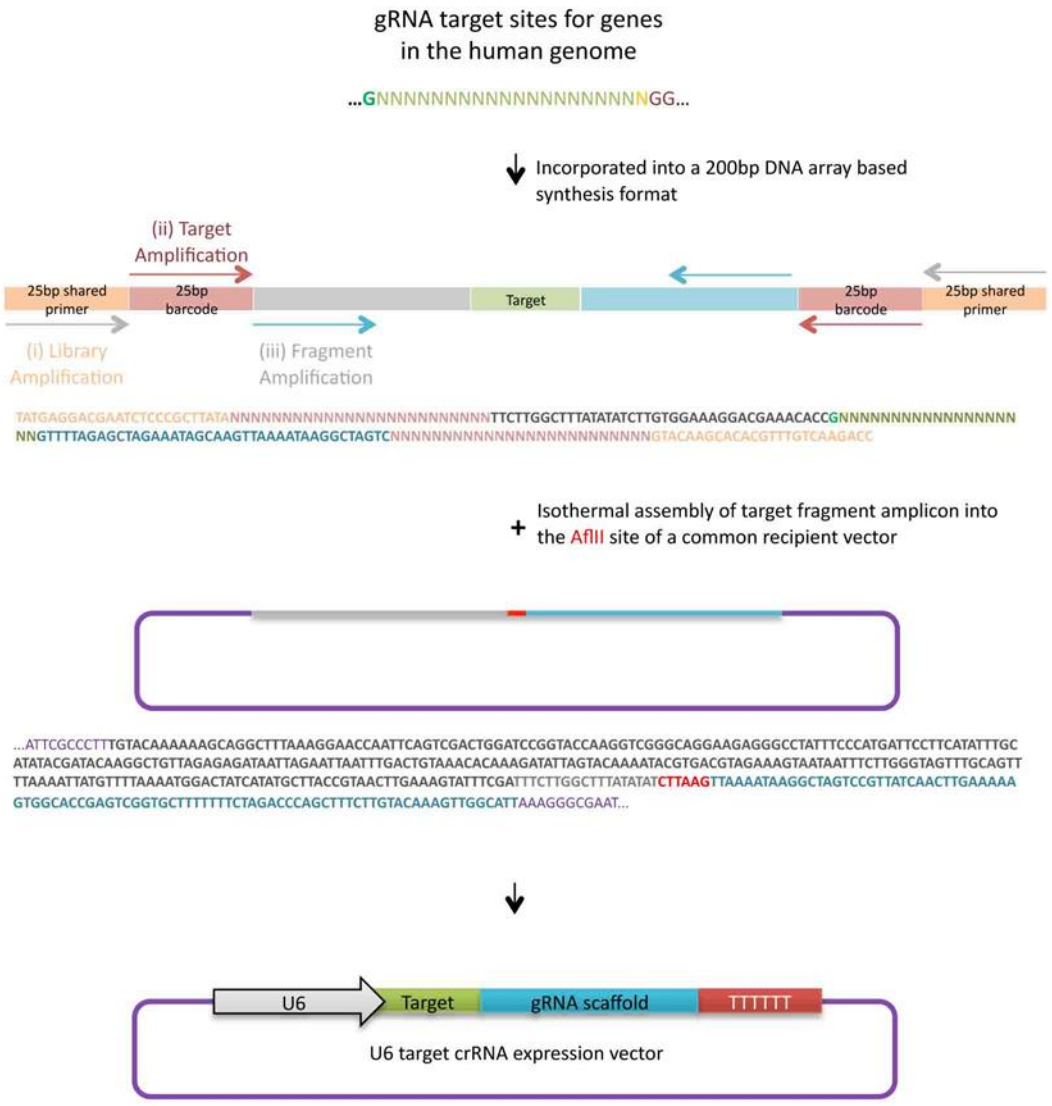


Figure 3_14. Methodology for multiplex synthesis, retrieval and U6 expression vector cloning of guide RNAs targeting genes in the human genome.

We established a resource of ~190k bioinformatically computed unique gRNA sites targeting ~40.5% of all exons of genes in the human genome (list in Supplementary Table 1). Furthermore, we incorporated these into a 200bp format (list in Supplementary Table 2) that is compatible for multiplex synthesis on DNA arrays. Specifically, our design allows for (i) targeted retrieval of a specific or pools of gRNA targets from the DNA array oligonucleotide pool (through 3 sequential rounds of nested PCR as indicated in the figure schematic); and (ii) its rapid cloning into a common expression vector which upon linearization using an AflII site serves as a recipient for Gibson assembly mediated incorporation of the gRNA insert fragment.

Figure 3_15. Continued) Figure 3_15. Evaluating the landscape of targeting by Cas9-gRNA complexes and TALEs.
(A) The methodology of our approach is outlined:
(B) Construct libraries are generated with a biased distribution of binding site sequences and random sequence 24bp tags that will be incorporated into reporter gene transcripts (top). The transcribed tags are highly degenerate so that they should map many-to-one to Cas9 or TALE binding sequences. The construct libraries are sequenced (3^{rd} level, left) to establish which tags co-occur with binding sites, resulting in an association table of binding sites *vs.* transcribed tags (4^{th} level, left). Multiple construct libraries built for different binding sites may be sequenced at once using library barcodes (indicated here by the light blue and light yellow colors; levels 1-4, left). A construct library is then transfected into a cell population and a set of different Cas9/gRNA or TALE transcription factors are induced in samples of the populations (2^{nd} level, right). One sample is always induced with a fixed TALE activator targeted to a fixed binding site sequence within the construct (top level, green box); this sample serves as a positive control (green sample, also indicated by a + sign). cDNAs generated from the reporter mRNA molecules in the induced samples are then sequenced and analyzed to obtain tag counts for each tag in a sample (3^{rd} and 4^{th} level, right). As with the construct library sequencing, multiple samples, including the positive control, are sequenced and analyzed together by appending sample barcodes. Here the light red color indicates one non-control sample that has been sequenced and analyzed with the positive control (green). Because only the transcribed tags and not the construct binding sites appear in each read, the binding site *vs.* tag association table obtained from construct library sequencing is then used to tally up total counts of tags expressed from each binding site in each sample (5^{th} level). The tallies for each non-positive control sample are then converted to normalized expression levels for each binding site by dividing them by the tallies obtained in the positive control sample.
(C) The targeting landscape of a Cas9-gRNA complex reveals that it is on average tolerant to 1-3 mutations in its target sequences.
(D) The Cas9-gRNA complex is also largely insensitive to point mutations, except those localized to the PAM sequence. Notably this data reveals that the predicted PAM for the S. pyogenes Cas9 is not just NGG but also NAG.
(E) Introduction of 2 base mismatches significantly impairs the Cas9-gRNA complex activity, however only when these are localized to the 8-10 bases nearer the 3' end of the gRNA target sequence (in the heat plot the target sequence positions are labeled from 1-23 starting from the 5' end).
(F) Similarly examining the TALE off-targeting data for an 18-mer TALE reveals that it can tolerate on average 1-2 mutations in its target sequence, and fails to activate a large majority of 3 base mismatch variants in its targets.
(G) The 18-mer TALE is, similar to the Cas9-gRNA complexes, largely insensitive to single base mismatched in its target. Introduction of 2 base mismatches significantly impairs the 18-mer TALE activity. Notably we observe that TALE activity is more sensitive to mismatches nearer the 5' end of its target sequence (in the heat plot the target sequence positions are labeled from 1-18 starting from the 5' end).

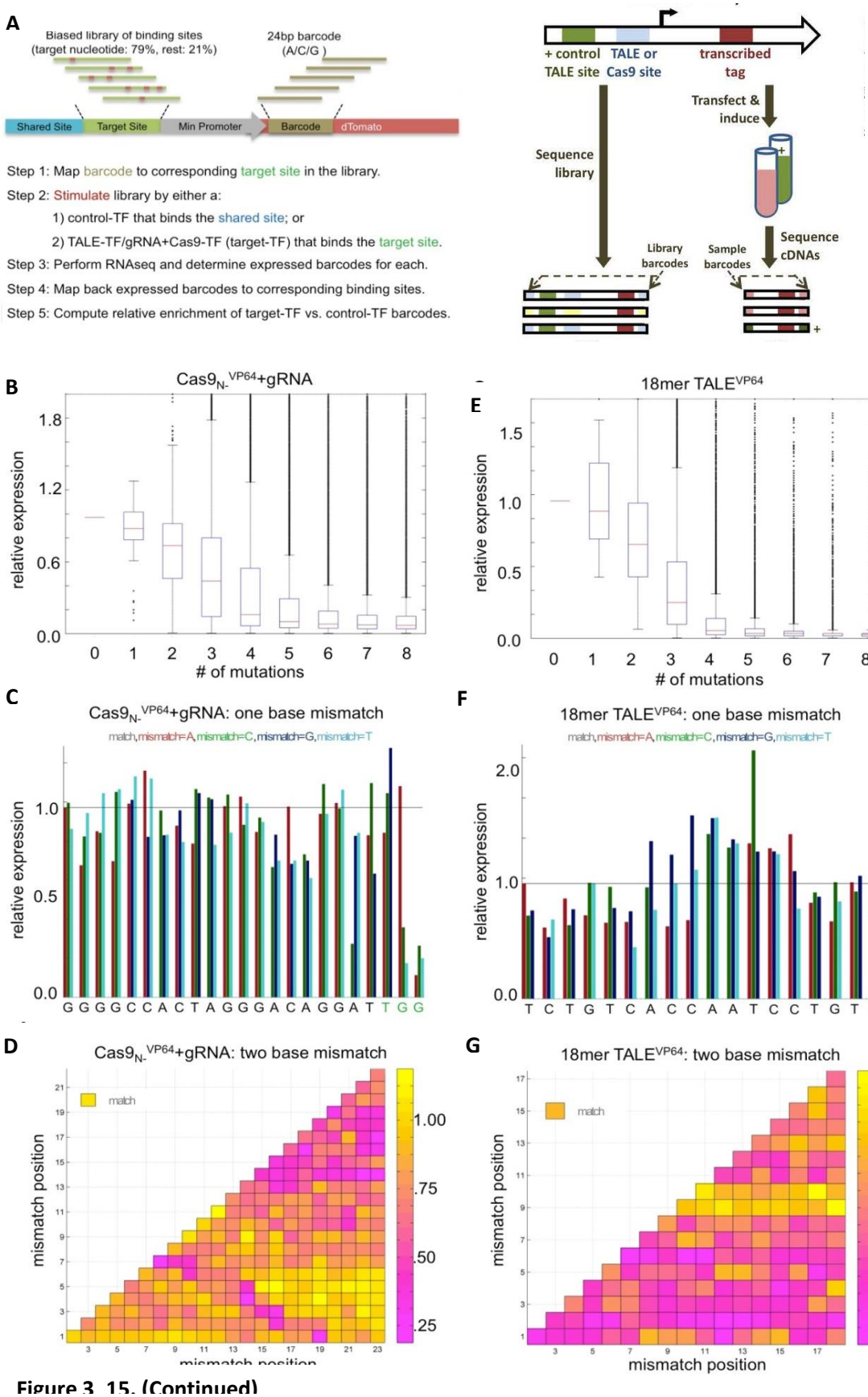


Figure 3. 15. (Continued)

we design a construct library in which each element of the library comprises a minimal promoter driving a dTomato fluorescent protein. Downstream of the transcription start site a 24bp (A/C/G) random transcript tag is inserted, while two TF binding sites are placed upstream of the promoter: one is a constant DNA sequence shared by all library elements, and the second is a variable feature that bears a 'biased' library of binding sites which are engineered to span a huge collection of sequences that present many combinations of mutations away from the target sequence the programmable DNA targeting complex was designed to bind. We achieved this using degenerate oligonucleotides engineered to bear nucleotide frequencies at each position such that the target sequence nucleotide appears at a 79% frequency and each other nucleotide occurs at 7% frequency. The reporter library is then sequenced to reveal the associations between the 24bp dTomato transcript tags and their corresponding 'biased' target site in the library element. The huge diversity of the transcript tags assures that sharing of tags between different targets will be extremely rare, while the biased construction of the target sequences means that sites with few mutations will be associated with more tags than sites with more mutations. Next we stimulate transcription of the dTomato reporter genes with either a control-TF engineered to bind the shared DNA site, or the target-TF that was engineered to bind the target site. As assayed by dTomato fluorescence, protein expression was observed to peak by ~ 48 hours and thus to prevent over-stimulation of the library total RNA was harvested within 24 hours. We then measure the abundance of each expressed transcript tag in each sample by conducting RNAseq on the stimulated cells, and then map these back to their corresponding binding sites using the association table established earlier. Note that one would expect the control-TF to excite all library members equally since its binding site is shared across all library elements, while the target-TF will skew the distribution of the expressed members to those that are preferentially

targeted by it. This assumption is used in step 5 to compute a normalized expression level for each binding site by dividing the tag counts obtained for the target-TF by those obtained for the control-TF.

We used the above approach to analyze the targeting landscape of multiple Cas9-gRNA complexes. These complexes on average tolerate 1-3 mutations in their target sequences (Figure 3_15 B). They are also largely insensitive to point mutations, except those localized to the PAM sequence (Figure 3_15 C). Introduction of 2 base mismatches significantly impairs activity, but only when these are localized to the 8-10 bases nearer the 3' end of the gRNA target sequence (Figure 3_15 D). These results are further reaffirmed by specificity data generated using two different Cas9-gRNA complexes (Figure 3_16). Notably we found that different gRNAs can have vastly different specificity profiles (Figure 3_16 A, B), specifically, gRNA2 here tolerates up to 3 mismatches and gRNA3 only up to 1. We next ran an array of experiments to validate these results. We also confirmed via targeted experiments that single-base mismatches within 12bp of the 3' end of the spacer in the assayed gRNAs indeed still result in detectable targeting, however 2bp mismatches in this region result in rapid loss of activity (Figure 3_17). An interesting aspect of the single-base mismatch data from both these experiments was that the predicted PAM for the *S. pyogenes* Cas9 is not just NGG but also NAG[20]. We confirmed this result with targeted experiments using the wild-type Cas9 in a nuclease assay (Figure 3_18). Taken together, our data demonstrate that the Cas9-

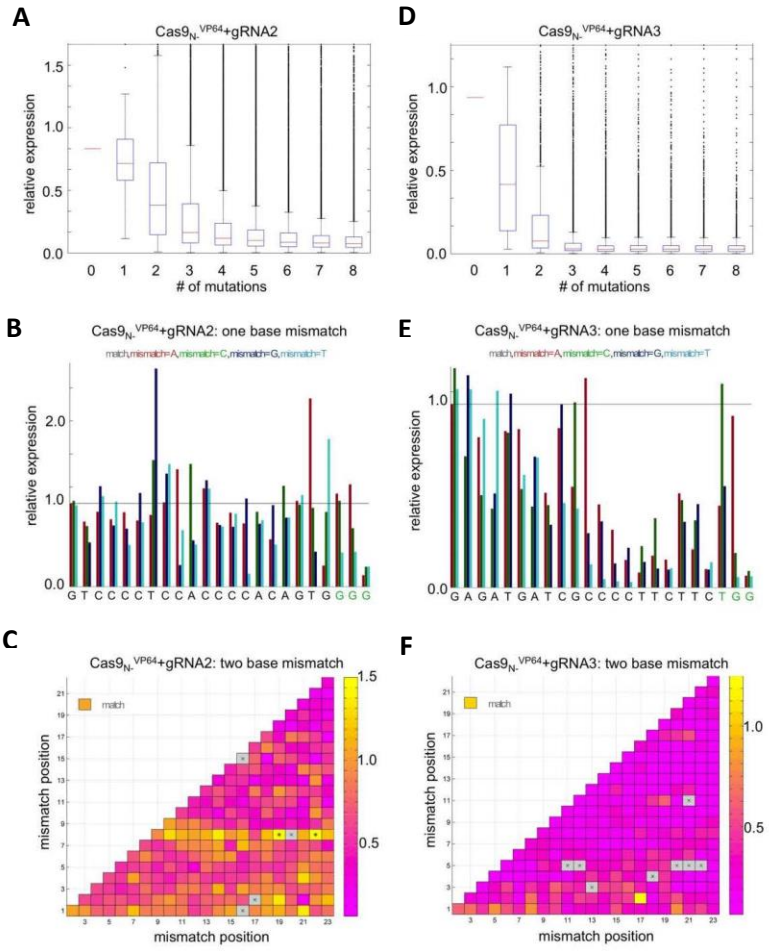


Figure 3_16. Evaluating the landscape of targeting by Cas9-gRNA complexes.
Using the approach described in Figure 3_15 we analyzed the targeting landscape of two additional Cas9-gRNA complexes (A-C) and (D-F). Notably we find that these two gRNAs have vastly different specificity profiles with gRNA2 tolerating up to 2-3 mismatches and gRNA3 only up to 1. These aspects are reflected in both the one base mismatch (B,E) and two base mismatch plots (C,F). To improve display, data outliers highlighted by 'x' symbols were not displayed in (C,F).

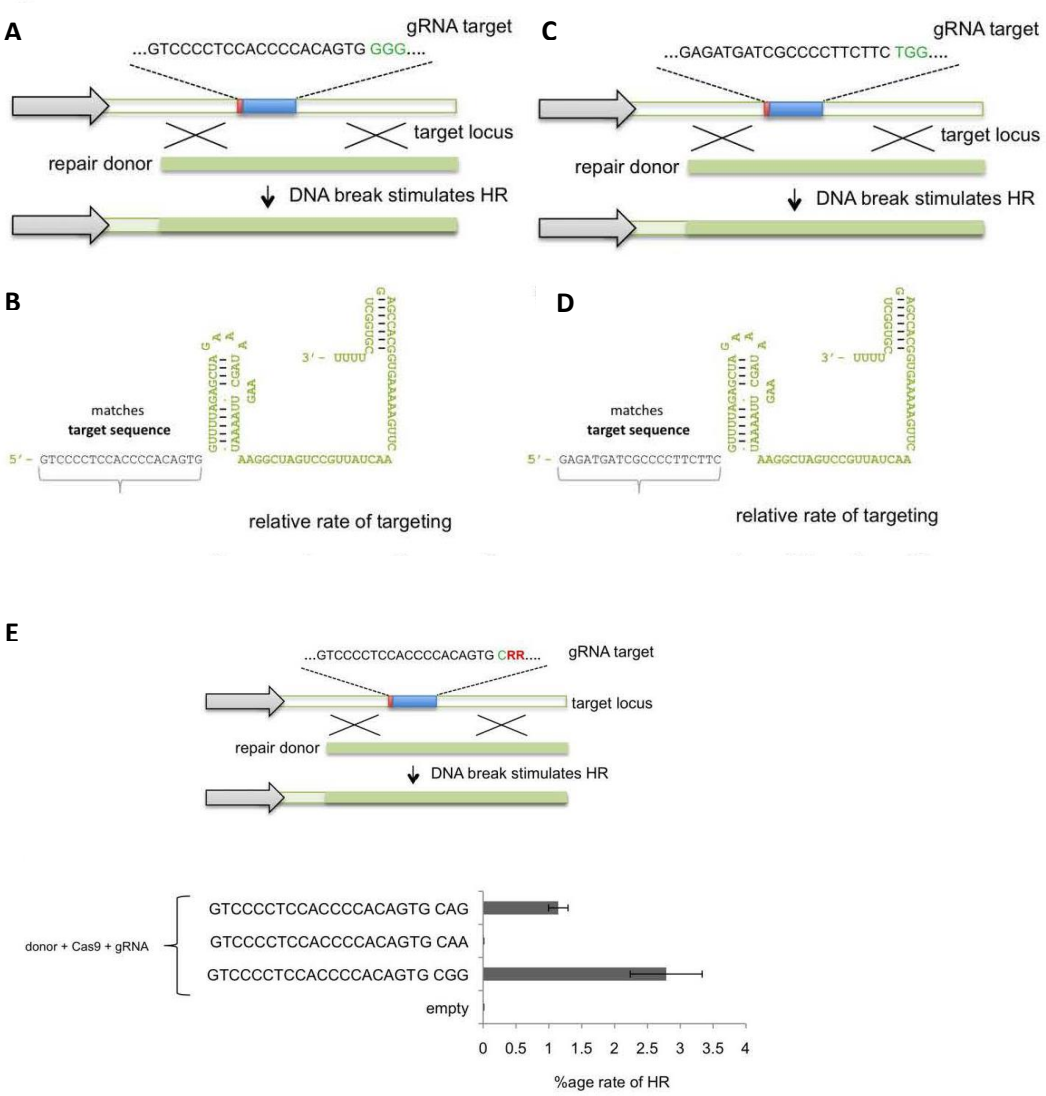


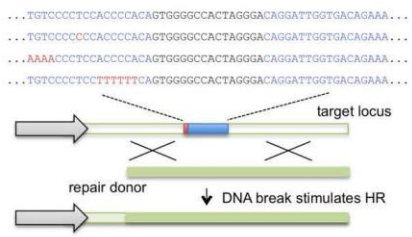
Figure 3_17 Validations, single and double-base gRNA mismatches. Using a nuclease assay we tested 2 independent gRNAs: gRNA2 (A,B) and gRNA3 (C,D) bearing single or double-base mismatches (highlighted in red) in the spacer sequence versus the target. These experiments confirmed that single-base mismatches within 12bp of the 3' end of the spacer in the assayed gRNAs indeed still result in detectable targeting, however 2bp mismatches in this regions result in rapid loss of activity. These results further highlight the differences in specificity profiles between different gRNAs consistent with the results in Figure 3_16

gRNA system can tolerate multiple mismatches in its target sequence. Consequently, achieving high targeting specificity with current experimental formats will likely require judicious and potentially complicated bioinformatic choice of gRNAs. Indeed, when we rescanned a previously generated set of ~190K Cas9 targets in human exons that had no alternate NGG targets sharing the last 13nt of the targeting sequence for the absence of alternate NGG and NAG sites at least one mismatch away, only .04% were specific at this level.

We next applied our transcriptional specificity assay to examine the mutational tolerance of another widely used genome editing tool, TALE domains. Examining the TALE off-targeting data (Figure 3_15) reveals that 18-mer TALEs tolerate 1-2 mutations in their target sequences, but fail to activate a large majority of 3 base mismatch variants in their targets. They are also particularly sensitive to mismatches nearer the 5' end of their target sequences. Intriguingly certain mutations in the middle of the target lead to higher TALE activity, an aspect that needs further evaluation. We also observed that shorter TALEs (14-mer and 10-mer) are progressively more specific in their targeting but also reduced in activity by nearly an order of magnitude (Figure 3_18). Taken together, these data imply that engineering shorter TALEs or TALEs bearing a judicious composition of high and low affinity monomers can potentially yield higher specificity in genome engineering applications, while the requirement for FokI dimerization in nuclease applications is essential to avoid off-target effects for the shorter TALEs (*20–22*).

Unlike TALEs where direct control of the TALE size or monomer composition is a ready approach to modulating specificity, there are limited current avenues for

A



B

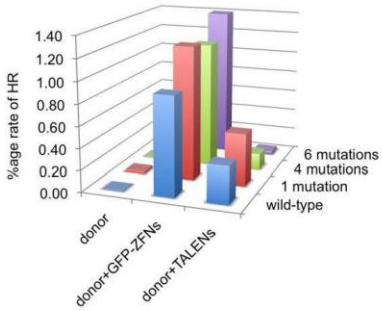


Figure 3_18. Validations, TALE mutations.
Using a nuclease mediated HR assay (A,B) we confirmed that 18-mer TALEs indeed tolerate multiple mutations in their target sequences.

engineering the Cas9-gRNA complex towards lower binding affinity (and hence higher specificity) for their targets (*23*, *24*). We therefore focused on exploiting cooperativity requirements to improve specificity. In the context of genome-editing, we chose to focus on creating off-set nicks. Our motivation stems from the observation that a large majority of nicks seldom result in NHEJ events (*25*), thus minimizing the effects of off-target nicking. Towards this we found that inducing off-set nicks to generate DSBs is highly effective at inducing gene disruption at both integrated reporter constructs and at native genomic loci (Figure 3_19). Interestingly we noted that consistent with the standard model for HR mediated repair (*26*) engineering of 5' overhangs via off-set nicks generated more robust NHEJ events as opposed to 3' overhangs (Figure 3_19 B). Intriguingly generation of 3' overhangs did not result in improvement of HR rates (Figure 3_19 C). It remains to be determined if Cas9 biochemistry or chromatin state and nucleotide composition of the genomic loci also contributed to the observed asymmetry in targeting rates at the two loci tested above. Taken together, we conclude that use of nicks for HR and off-set nicks for generating DSBs offers a promising route for mitigating the effects of off-target Cas9-gRNA activity.

Discussion

Our results demonstrate the promise of CRISPR-mediated gene targeting for RNA-guided, robust, and multiplexable mammalian genome engineering. The ease of retargeting our system to modify genomic sequences greatly exceeds that of comparable ZFNs and TALENs, while offering similar or greater efficiencies (*27*). Existing studies of type II CRISPR specificity efficiencies (*27*) suggest that target sites must perfectly match

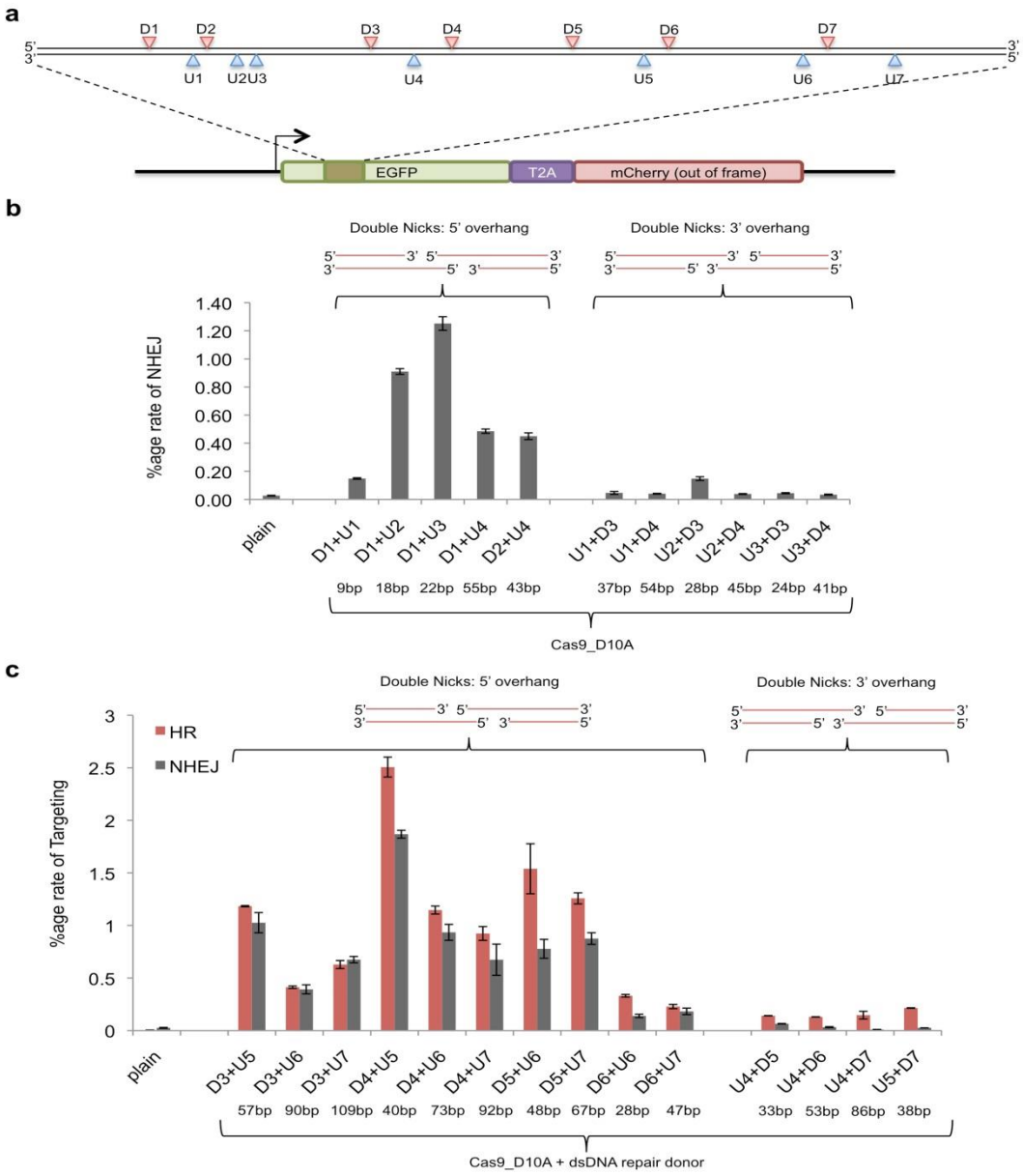


Figure 3_19. off-set nicking

(A) We employed the traffic light reporter to simultaneously assay for HR and NHEJ events upon introduction of targeted nicks or breaks: DNA cleavage events resolved through the HDR pathway restore the GFP sequence, whereas mutagenic NHEJ causes frame-shifts rendering the GFP out of frame and the downstream mCherry sequence in frame. For the assay, we designed 14 gRNAs covering a 200bp stretch of DNA: 7 targeting the sense strand (U1-7) and 7 the antisense strand (D1-7). Using the Cas9D10A mutant, which nicks the complementary strand, we used different two-way combinations of the gRNAs to induce a range of programmed 5' or 3' overhangs (the nicking sites for the 14 gRNAs are indicated).

(B) Inducing off-set nicks to generate DSBs is highly effective at inducing gene disruption. Notably off-set nicks leading to 5' overhangs result in more NHEJ events as opposed to 3' overhangs.

(C) Similarly, generating 3' overhangs also favors the ratio of HR over NHEJ events, but the total number of HR events is significantly lower than when a 5' overhang is generated. In (b,c) the predicted overhang lengths are indicated below the corresponding x-axis legends.

the PAM sequence NGG and the 8- to 12-base "seed sequence" at the 3′ end of the gRNA. The importance of the remaining 8 to 12 bases is less well understood and may depend on the binding strength of the matching gRNAs or on the inherent tolerance of Cas9 itself. Indeed, Cas9 will tolerate single mismatches at the 5′ end in bacteria and in vitro, which suggests that the 5′ G is not required. Moreover, it is likely that the target locus's underlying chromatin structure and epigenetic state will also affect the efficiency of genome editing in eukaryotic cells (*16*), although we suspect that Cas9's helicase activity may render it more robust to these factors, but this remains to be evaluated. In addition, the range of CRISPR-targetable sequences could be expanded through the use of homologs with different PAM requirements (*28*) or by directed evolution. Finally, inactivating one of the Cas9 nuclease domains increases the ratio of HR to NHEJ and may reduce toxicity (*29, 30*), whereas inactivating both domains may enableCas9 to function as a retargetable DNA binding protein.

To illustrate and improve the specific of genome targeting tools is uppermost important for its application in biomedical research and gene therapy. Here, we observed that the Ca9-gRNA system can result in significant off-targeting events. Interestingly we note that there are huge differences in specificity between evaluated gRNAs. Based on this we speculate that likely the Cas9 protein contributes primarily to PAM recognition, but gRNA-DNA binding (and associated thermodynamic parameters) are a prominent determinant of specificity. Thus judicious choice of gRNAs will be a productive route to improved target specificity, albeit rules governing their design such as Tm, nucleotide composition, secondary structure of gRNA spacer versus scaffold, and role of underlying chromatin structure of the target loci remain to be determined. Controlling the dose and duration of Cas9 and gRNA expression will also be critical for engineering high specificity, and thus RNA based delivery will be an attractive genome

editing route (*31*, *32*). While structure-guided design and directed evolution may eventually improve the specificity of individual Cas9 proteins, we have also shown here that engineering a requirement for cooperativity via off-set nicking to generate DSBs can potentially ameliorate off-target activity, and may be an important avenue for exploring therapeutic applications. The improved ease and efficacy of editing and regulating genomes using this RNA-guided genome engineering approach will have broad implications for our ability to tune and program complex biological systems.

With enhanced activity and specificity, we expect that RNA- guided genome targeting will have broad implications for synthetic biology (*13*, *33*), the direct and multiplexed perturbation of gene networks (*16*, *34*), and targeted ex vivo (*35*, *36*) and in vivo gene therapy (*37*).

Materials and Methods

Plasmid construction

The Cas9 gene sequence was human codon optimized and assembled by hierarchical fusion PCR assembly of 9 500bp gBlocks ordered from IDT (sequence in fig. S1A). Cas9_D10A was similarly constructed. The resulting full-length products were cloned into the pcDNA3.3-TOPO vector (Invitrogen). The target gRNA expression constructs were directly ordered as individual 455bp gBlocks from IDT and either cloned into the pCR-BluntII-TOPO vector (Invitrogen) or pcr amplified. The vectors for the HR reporter assay involving a broken GFP were constructed by fusion PCR assembly of the GFP sequence bearing the stop codon and 68bp AAVS1 fragment, or 58bp fragments from the DNMT3a and DNMT3b genomic loci assembled

into the EGIP lentivector from Addgene (plasmid #26777). These lentivectors were then used to establish the GFP reporter stable lines. TALENs used in this study were constructed using the protocols described in (*14*). All DNA reagents developed in this study are available at Addgene.

Cell culture

PGP1 iPS cells were maintained on Matrigel (BD Biosciences)-coated plates in mTeSR1 (Stemcell Technologies). Cultures were passaged every 5–7 d with TrypLE Express (Invitrogen). K562 cells were grown and maintained in RPMI (Invitrogen) containing 15% FBS. HEK 293T cells were cultured in Dulbecco's modified Eagle's medium (DMEM, Invitrogen) high glucose supplemented with 10% fetal bovine serum (FBS, Invitrogen), penicillin/streptomycin (pen/strep, Invitrogen), and non-essential amino acids (NEAA, Invitrogen). All cells were maintained at 37°C and 5% CO_2 in a humidified incubator.

Gene targeting of PGP1 iPS, K562 and 293Ts

PGP1 iPS cells were cultured in Rho kinase (ROCK) inhibitor (Calbiochem) 2h before nucleofection. Cells were harvest using TrypLE Express (Invitrogen) and 2×10^6 cells were resuspended in P3 reagent (Lonza) with 1μg Cas9 plasmid, 1μg gRNA and/or 1μg DNA donor plasmid, and nucleofected according to manufacturer's instruction (Lonza). Cells were subsequently plated on an mTeSR1-coated plate in mTeSR1 medium supplemented with ROCK inhibitor for the first 24h. For K562s, 2×10^6 cells were resuspended in SF reagent (Lonza) with 1μg Cas9 plasmid, 1μg gRNA and/or 1μg DNA donor plasmid, and nucleofected according to manufacturer's instruction (Lonza). For 293Ts, 0.1×10^6 cells were transfected with 1μg Cas9 plasmid, 1μg gRNA and/or 1μg DNA donor plasmid using Lipofectamine 2000 as per the

manufacturer's protocols. The DNA donors used for endogenous AAVS1 targeting were either a dsDNA donor (Figure 3_2) or a 90mer oligonucleotide. The former has flanking short homology arms and a SA-2A-puromycin-CaGGS-eGFP cassette to enrich for successfully targeted cells.

Assess the targeting efficiency

Cells were harvested 3 days after nucleofection and the genomic DNA of ~1 X 10^6 cells was extracted using prepGEM (ZyGEM). PCR was conducted to amplify the targeting region with genomic DNA derived from the cells and amplicons were deep sequenced by MiSeq Personal Sequencer (Illumina) with coverage >200,000 reads. The sequencing data was analyzed to estimate NHEJ efficiencies. The reference AAVS1 sequence analyzed is:

```
CACTTCAGGACAGCATGTTTGCTGCCTCCAGGGATCCTGTGTCCCCGAGCTGGGACCACCTTATATTCCCAGGGCCG
GTTAATGTGGCTCTGGTTCTGGGTACTTTTATCTGTCCCCTCCACCCCACAGTGGGGCCACTAGGGACAGGATTGGT
GACAGAAAAGCCCCATCCTTAGGCCTCCTCCTTCCTAGTCTCCTGATATTGGGTCTAACCCCCACCTCCTGTTAGGC
AGATTCCTTATCTGGTGACACACCCCCATTTCCTGGA
```

The PCR primers for amplifying the targeting regions in the human genome are:

AAVS1-R CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTacaggaggtgggggttagac

AAVS1-F.1 ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGATtatattcccagggccggtta

AAVS1-F.2 ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATCGtatattcccagggccggtta

AAVS1-F.3 ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCTAAtatattcccagggccggtta

AAVS1-F.4 ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGTCAtatattcccagggccggtta

AAVS1-F.5 ACACTCTTTCCCTACACGACGCTCTTCCGATCTCACTGTtatattcccagggccggtta

AAVS1-F.6 ACACTCTTTCCCTACACGACGCTCTTCCGATCTATTGGCtatattcccagggccggtta

AAVS1-F.7 ACACTCTTTCCCTACACGACGCTCTTCCGATCTGATCTGtatattcccagggccggtta

AAVS1-F.8 ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCAAGTtatattcccagggccggtta

AAVS1-F.9 ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGATCtatattcccagggccggtta

AAVS1-F.10 ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGCTAtatattcccagggccggtta

AAVS1-F.11 ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTAGCCtatattcccagggccggtta

AAVS1-F.12 ACACTCTTTCCCTACACGACGCTCTTCCGATCTTACAAGtatattcccagggccggtta

To analyze the HR events using the DNA donor in Fig. 2C the primers used were:

HR_AAVS1-F CTGCCGTCTCTCTCCTGAGT

HR_Puro-R GTGGGCTTGTACTCGGTCAT

Bioinformatics approach for computing human exon CRISPR targets and methodology for their multiplexed synthesis

We sought to generate a set of gRNA gene sequences that maximally target specific locations in human exons but minimally target other locations in the genome. Maximally efficient targeting by a gRNA is achieved by 23nt sequences, the 5'-most 20nt of which exactly complement a desired location, while the three 3'-most bases must be of the form NGG. Additionally, the 5'-most nt must be a G to establish a pol-III transcription start site. However, according to (*38*), mispairing of the six 5'-most nt of a 20bp gRNA against its genomic target does not abrogate Cas9-mediated cleavage so long as the last 14nt pairs properly, but mispairing of the eight 5'-most nt along with pairing of the last 12 nt does, while the case of the seven 5-most nt mispairs and 13 3' pairs was not tested. To be conservative regarding off-target effects, we therefore assumed that the case of the seven 5'-most mispairs is, like the case of six, permissive of cleavage, so that pairing of the 3'-most 13nt is sufficient for cleavage. To identify CRISPR target sites within human exons that should be cleavable without off-target cuts, we therefore examined all 23bp sequences of the form 5'-GBBBB BBBBB BBBBB BBBBB NGG-3' (*form 1*), where the B's represent the bases at the exon location, for which no sequence of the

form 5'-NNNNN NNBBB BBBBB BBBBB NGG-3' (*form 2*) existed at any other location in the human genome. Specifically, we (i) downloaded a BED file of locations of coding regions of all RefSeq genes the GRCh37/hg19 human genome from the UCSC Genome Browser (*39–41*). Coding exon locations in this BED file comprised a set of 346089 mappings of RefSeq mRNA accessions to the hg19 genome. However, some RefSeq mRNA accessions mapped to multiple genomic locations (probable gene duplications), and many accessions mapped to subsets of the same set of exon locations (multiple isoforms of the same genes). To distinguish apparently duplicated gene instances and consolidate multiple references to the same genomic exon instance by multiple RefSeq isoform accessions, we therefore (ii) added unique numerical suffixes to 705 RefSeq accession numbers that had multiple genomic locations, and (iii) used the mergeBed function of BEDTools (*42*) (v2.16.2-zip-87e3926) to consolidate overlapping exon locations into merged exon regions. These steps reduced the initial set of 346089 RefSeq exon locations to 192783 distinct genomic regions. We then downloaded the hg19 sequence for all merged exon regions using the UCSC Table Browser, adding 20bp of padding on each end. (iv) Using custom perl code, we identified 1657793 instances of *form 1* within this exonic sequence. (v) We then filtered these sequences for the existence of off-target occurrences of *form* 2: For each merged exon *form 1* target, we extracted the 3'-most 13bp specific (B) "core" sequences and, for each core generated the four 16bp sequences 5'-BBB BBBBB BBBBB NGG-3' (N = A, C, G, and T), and searched the entire hg19 genome for exact matches to these 6631172 sequences using Bowtie version 0.12.8 (*43*) using the parameters -l 16 -v 0 -k 2. We rejected any exon target site for which there was more than a single match. Note that because any specific 13bp core sequence followed by the sequence NGG confers only 15bp of specificity, there should be on average ~5.6 matches to an extended core sequence in a random ~3Gb sequence (both strands).

Therefore, most of the 1657793 initially identified targets were rejected; however 189864 sequences passed this filter. These comprise our set of CRISPR-targetable exonic locations in the human genome. The 189864 sequences target locations in 78028 merged exonic regions (~40.5% of the total of 192783 merged human exon regions) at a multiplicity of ~2.4 sites per targeted exonic region. To assess targeting at a gene level, we clustered RefSeq mRNA mappings so that any two RefSeq accessions (including the gene duplicates we distinguished in (ii)) that overlap a merged exon region are counted as a single gene cluster, the 189864 exonic specific CRISPR sites target 17104 out of 18872 gene clusters (~90.6% of all gene clusters) at a multiplicity of ~11.1 per targeted gene cluster. (Note that while these gene clusters collapse RefSeq mRNA accessions that represent multiple isoforms of a single transcribed gene into a single entity, they will also collapse overlapping distinct genes as well as genes with antisense transcripts.) At the level of original RefSeq accessions, the 189864 sequences targeted exonic regions in 30563 out of a total of 43726 (~69.9%) mapped RefSeq accessions (including our distinguished gene duplicates) at a multiplicity of ~6.2 sites per targeted mapped RefSeq accession.

As we gather information on CRISPR performance at our computationally predicted human exon CRISPR target sites, we plan to refine our database by correlating performance with factors we expect to be important, such as base composition and secondary structure of both gRNAs and genomic targets (*44*), and the epigenetic state of these targets in human cell lines for which this information is available.

Finally, we also incorporated these target sequences into a 200bp format that is compatible for multiplex synthesis on DNA arrays. Our design allows for targeted retrieval of a specific or pools of gRNA sequences from the DNA array based oligonucleotide pool and its rapid cloning into a common expression vector. Specifically we tested this approach by

synthesizing a 12k oligonucleotide pool from CustomArray Inc. Furthermore, as per our approach we were able to successfully retrieve gRNAs of choice from this library (Figure 3_14). We observed an error rate of ~4 mutations per 1000bp of synthesized DNA.

Cell culture and transfections

HEK 293T cells were cultured in Dulbecco's modified Eagle's medium (DMEM, Invitrogen) high glucose supplemented with 10% fetal bovine serum (FBS, Invitrogen), penicillin/streptomycin (pen/strep, Invitrogen), and non-essential amino acids (NEAA, Invitrogen). Cells were maintained at 37°C and 5% CO_2 in a humidified incubator.

Transfections involving nuclease assays were as follows: 0.4×10^6 cells were transfected with 2μg Cas9 plasmid, 2μg gRNA and/or 2μg DNA donor plasmid using Lipofectamine 2000 as per the manufacturer's protocols. Cells were harvested 3 days after transfection and either analyzed by FACS, or for direct assay of genomic cuts the genomic DNA of ~1 X 10^6 cells was extracted using DNAeasy kit (Qiagen). For these PCR was conducted to amplify the targeting region with genomic DNA derived from the cells and amplicons were deep sequenced by MiSeq Personal Sequencer (Illumina) with coverage >200,000 reads. The sequencing data was analyzed to estimate NHEJ efficiencies.

For transfections involving transcriptional activation assays: 0.4×10^6 cells were transfected with (1) 2μg $Cas9_{N-}$VP64 plasmid, 2μg gRNA and/or 0.25μg of reporter construct; or (2) 2μg $Cas9_{N-}$ plasmid, 2μg MS2-VP64, 2μg gRNA-2XMS2aptamer and/or 0.25μg of reporter construct. Cells were harvested 24-48hrs post transfection and assayed using FACS or immunofluorescence methods, or their total RNA was extracted and these were subsequently

analyzed by RT-PCR. Here standard taqman probes from Invitrogen for REX1, OCT4, SOX2 and NANOG were used, with normalization for each sample performed against GAPDH.

For transfections involving transcriptional activation assays for specificity profile of Cas9-gRNA complexes and TALEs: 0.4×10^6 cells were transfected with (1) 2μg $Cas9_N$-VP64 plasmid, 2μg gRNA and 0.25μg of reporter library; or (2) 2μg TALE-TF plasmid and 0.25μg of reporter library; or (3) 2μg control-TF plasmid and 0.25μg of reporter library. Cells were harvested 24hrs post transfection (to avoid the stimulation of reporters being in saturation mode). Total RNA extraction was performed using RNAeasy-plus kit (Qiagen), and standard RT-pcr performed using Superscript-III (Invitrogen). Libraries for next-generation sequencing were generated by targeted pcr amplification of the transcript-tags.

References

1. K. S. Makarova *et al.*, Evolution and classification of the CRISPR-Cas systems., *Nature reviews. Microbiology* **9**, 467–77 (2011).

2. M. Jinek *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity., *Science (New York, N.Y.)* **337**, 816–21 (2012).

3. P. Horvath, R. Barrangou, CRISPR/Cas, the immune system of bacteria and archaea., *Science (New York, N.Y.)* **327**, 167–70 (2010).

4. H. Deveau *et al.*, Phage response to CRISPR-encoded resistance in Streptococcus thermophilus., *Journal of bacteriology* **190**, 1390–400 (2008).

5. J. R. van der Ploeg, Analysis of CRISPR in Streptococcus mutans suggests frequent occurrence of acquired immunity against infection by M102-like bacteriophages., *Microbiology (Reading, England)* **155**, 1966–76 (2009).

6. M. Rho, Y.-W. Wu, H. Tang, T. G. Doak, Y. Ye, Diverse CRISPRs evolving in human microbiomes., *PLoS genetics* **8**, e1002441 (2012).

7. D. T. Pride *et al.*, Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time., *Genome research* **21**, 126–36 (2011).

8. G. Gasiunas, R. Barrangou, P. Horvath, V. Siksnys, Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria., *Proceedings of the National Academy of Sciences of the United States of America* **109**, E2579–86 (2012).

9. R. Sapranauskas *et al.*, The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli., *Nucleic acids research* **39**, 9275–82 (2011).

10. J. E. Garneau *et al.*, The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA., *Nature* **468**, 67–71 (2010).

11. R. Barrangou *et al.*, CRISPR provides acquired resistance against viruses in prokaryotes., *Science (New York, N.Y.)* **315**, 1709–12 (2007).

12. T. R. Brummelkamp, R. Bernards, R. Agami, A system for stable expression of short interfering RNAs in mammalian cells., *Science (New York, N.Y.)* **296**, 550–3 (2002).

13. J. Zou *et al.*, Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells., *Cell stem cell* **5**, 97–110 (2009).

14. N. E. Sanjana *et al.*, A transcription activator-like effector toolbox for genome engineering., *Nature protocols* **7**, 171–92 (2012).

15. J.-H. Lee *et al.*, A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells., *PLoS genetics* **5**, e1000718 (2009).

16. D. Hockemeyer *et al.*, Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases., *Nature biotechnology* **27**, 851–7 (2009).

17. A. P. Boyle *et al.*, High-resolution mapping and characterization of open chromatin across the genome., *Cell* **132**, 311–22 (2008).

18. S. Kosuri *et al.*, Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips., *Nature biotechnology* **28**, 1295–9 (2010).

19. W. Jiang, D. Bikard, D. Cox, F. Zhang, L. a Marraffini, RNA-guided editing of bacterial genomes using CRISPR-Cas systems., *Nature biotechnology* **31**, 233–239 (2013).

20. M. H. Porteus, J. P. Connelly, S. M. Pruett, A Look to Future Directions in Gene Therapy Research for Monogenic Diseases, **2** (2006), doi:10.1371/journal.pgen.0020133.

21. J. F. Meckler *et al.*, Quantitative analysis of TALE-DNA interactions suggests polarity effects., *Nucleic acids research* **41**, 4118–28 (2013).

22. S. W. Cho, S. Kim, J. M. Kim, J.-S. Kim, Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease., *Nature biotechnology* **31**, 230–232 (2013).

23. V. Pattanayak, C. L. Ramirez, J. K. Joung, D. R. Liu, Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection., *Nature methods* **8**, 765–70 (2011).

24. R. Gabriel *et al.*, An unbiased genome-wide analysis of zinc-finger nuclease specificity., *Nature biotechnology* **29**, 816–23 (2011).

25. P. Mali *et al.*, RNA-guided human genome engineering via Cas9., *Science (New York, N.Y.)* **339**, 823–6 (2013).

26. L. S. Symington, J. Gautier, Double-strand break end resection and repair pathway choice., *Annual review of genetics* **45**, 247–71 (2011).

27. B. Wiedenheft, S. H. Sternberg, J. a Doudna, RNA-guided genetic silencing systems in bacteria and archaea., *Nature* **482**, 331–8 (2012).

28. E. Deltcheva *et al.*, CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III., *Nature* **471**, 602–7 (2011).

29. R. Sapranauskas *et al.*, The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli., *Nucleic acids research* **39**, 9275–82 (2011).

30. M. Jinek *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity., *Science (New York, N.Y.)* **337**, 816–21 (2012).

31. S. W. Cho, S. Kim, J. M. Kim, J.-S. Kim, Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease., *Nature biotechnology* **31**, 230–232 (2013).

32. Z. Yu *et al.*, Highly Efficient Genome Modifications Mediated by CRISPR/Cas9 in Drosophila., *Genetics* (2013), doi:10.1534/genetics.113.153825.

33. S. S. Ajay, S. C. J. Parker, H. O. Abaan, K. V. F. Fajardo, E. H. Margulies, Accurate and comprehensive sequencing of personal genomes., *Genome research* **21**, 1498–505 (2011).

34. D. Hockemeyer *et al.*, Genetic engineering of human pluripotent cells using TALE nucleases., *Nature biotechnology* **29**, 731–4 (2011).

35. F. D. Urnov, E. J. Rebar, M. C. Holmes, H. S. Zhang, P. D. Gregory, Genome editing with engineered zinc finger nucleases., *Nature reviews. Genetics* **11**, 636–46 (2010).

36. A. Lombardo *et al.*, Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery., *Nature biotechnology* **25**, 1298–306 (2007).

37. H. Li *et al.*, In vivo genome editing restores haemostasis in a mouse model of haemophilia., *Nature* **475**, 217–21 (2011).

38. D. Bhaya, M. Davison, R. Barrangou, CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation., *Annual review of genetics* **45**, 273–97 (2011).

39. W. J. Kent *et al.*, The Human Genome Browser at UCSC, *Genome Research* **12**, 996–1006 (2002).

40. D. Karolchik *et al.*, The UCSC Table Browser data retrieval tool., *Nucleic acids research* **32**, D493–6 (2004).

41. T. R. Dreszer *et al.*, The UCSC Genome Browser database: extensions and updates 2011., *Nucleic acids research* **40**, D918–23 (2012).

42. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features., *Bioinformatics (Oxford, England)* **26**, 841–2 (2010).

43. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome., *Genome biology* **10**, R25 (2009).

44. C. T. Kåhrström, Phage biology: Giving CRISPR the slip., *Nature reviews. Microbiology* **11**, 72 (2013).

Chapter 4

Genome editing with targeted deaminases

Luhan Yang[1,2±], Adrian W. Briggs[1]*, Wei Leong Chew[1,2]*, Prashant Mali[1], Marc Guell[1], John Aach[1], Daniel Bryan Goodman[1,3], David Cox[3], Venkataramanan Soundararajan[1], Feng Zhang[5], George Church[1,3,4,±]

[1] Department of Genetics, Harvard Medical School, Boston, MA
[2] Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA
[3] Harvard-MIT Division of Health Science and Technology, Cambridge, MA
[4] Wyss Institute for Biologically Inspired Engineering, Boston, MA
[5] Broad Institute of MIT and Harvard, Cambridge, MA
McGovern Institute for Brain Research, MIT, Cambridge, MA
Department of Brain and Cognitive Sciences, MIT Cambridge, MA
* Contributed equally to this work as joint second authors

Acknowledgement

This work was supported by National Human Genome Research Institute Center for Excellence in Genomics Science (P50 HG003170, G.M.C.). A.M.B and M.C.G. are supported by European Molecular Biology Organization (EMBO) Long Term Fellowships. W.L.C is supported by the Agency for Science, Technology and Research, Singapore. We thank S. Raman, X. Rios, S. Byrne, N. Eroshenko and H. Wang for critical readings of the manuscript. We thank K, Joung and his lab members (Harvard) for providing critical suggestions and advice. M.Wang and M. Neuberger for pTrc99A-AID (University of Cambridge, UK), J. Way (Wyss Institute) for pUC57-ZF, P. Huang (Harvard) for pROEX-HTa, F. Isaacs (Yale University) for pZE21G, and K, Temme and C, Voigt (University of California San Francisco) for pTac-T7polymerase.

Author contribution

L.Y. and G.M.C. conceived the study. L.Y. designed, performed, and analyzed the bacterial experiments. L.Y and A.W.B designed and performed the human cell experiments. W.L.C. contributed to the bacterial experiments and data analysis. P.M contributed to the human cell experiments. J.A contributed to the experimental design, data analysis and statistic tests. M.G.C and D.B.G. performed the whole-genome sequence data analysis. D.C contributed to the human cell experiments and data analysis. V.S. simulated AID structure to guide the protein design. F.Z. provided the TALE construct and provided critical advice on TALE-AID optimization. G.M.C. supervised the work of L.Y, A.W.B, W.L.C, P.M, M.G.C, A.J, D.B.G, V.S; G.M.C provided support for this study. L.Y. wrote the manuscript with support from all authors.

Summary

Tools for efficient and precise genome editing facilitate functional studies and advance gene therapies. Current nucleases-based methods that employ homologous recombination at sites of targeted double-strand breaks (DSBs) are limited by competing DNA repair pathways (*1*) and cytotoxicity (*2*). Here we present targeted cytidine deaminases, consisting of DNA deaminases (*3*) fused with programmable DNA-binding modules (*4, 5*), that perform sequence-specific genome editing without generating DSBs and the need to simultaneously provide replacement (i.e., donor) DNA. Targeted deaminases are both efficient and specific in Escherichia coli, converting a targeted cytidine to thymidine with 13% efficiency and 95% accuracy. Edited cells do not exhibit random hypermutation or aberrant genomic structural changes. These novel enzymes also function in human cells, causing a site-specific C:G->T:A transition in 2.5% of cells, with significantly less toxicity than nucleases. Targeted deaminases therefore represent a platform for safer and effective genome editing in prokaryotes and eukaryotes. The independence from DSBs and donor DNA suggests applications of this tool in multiplexed editing (including repetitive elements) and inducible genome editing in whole animals.

Introduction

Genome editing in mammalian cells has been greatly facilitated by the development of customized zinc finger (ZF)- and transcription activator-like effectors (TALEs)- (*4*) nucleases (ZFNs) (*5, 6*) and TALENs (*7, 8*) that create DSBs at specifically targeted sites in the genome. When exogenous donor DNA has been provided with arms homologous to the targeted sites, cells repair the DSB at high rates and with high precision through homologous recombination (HR) with the donor DNA. However, the use of targeted DSBs also imposes limitations; first,

during DSB repair, HR competes with non-homologous end joining (NHEJ) which does not require donor DNA and often introduces mutations at the repair site. In the absence of In the absence of nuclease-based methods, NHEJ occurs 30-fold to 40,000-fold more frequently than HR in human cells, so that effective use of targeted nucleases requires coordinating their expression with high levels of donor DNA (*1*). Second, DSBs are toxic to the cell and can introduce genome instability, placing further constraints on targeted nuclease expression. These conditions make it unlikely that that targeted nucleases can be used safely and effectively to make highly multiplexed changes to a genome, or to perform efficient gene targeting within multi-cellular organisms where delivery of donor DNA into cells at high copy numbers would be challenging. However, an accurate genome editing method that did not create DSBs or require donor DNA should escape these limitations.

Single-nucleotide genome editing independent of DSBs and DNA donor occurs naturally. Activation induced deaminase (AID) and apolipoprotein B mRNA editing enzyme catalytic polypeptide-like family proteins (APOBECs) (*3*) are cytidine deaminases expressed in vertebrates that act in the antibody diversification process or in the innate immune system as an agent that targets retroviruses (*3*). These enzymes can convert cytidines to uracils in DNA. If DNA replication occurs before uracil repair, the replication machinery will treat the uracil as thymine, leading to a C:G to T:A base pair conversion (*9*). This elegant editing mechanism suggests a potentially simple and effective genome editing approach that circumvents the limitations associated with nucleases. Here, we sought to test whether combining deaminases with DNA-binding proteins could target cytidine deamination to specific positions in the genome and thus enable targeted genome C:G→T:A editing.

Results

Design and functional test of targeted deaminase

As a first step, we engineered targeted deaminases by fusing the deaminases (APOBEC1, APOBEC3F, APOBEC3G (2K3A) (*10*) and AID) with a ZF recognizing the 9bp DNA sequence 5'-GCCGCAGTG-3' (*11*) (Figure 4_1 A). Based on the available structures of the deaminases (*12*), we inferred that the enzymes' domains reside at the C-terminus. We therefore tethered the ZF DNA-binding domains by a four amino-acid linker to the N-terminus of the deaminases to generate ZF-APOBEC1, 3F, 3G and ZF-AID (Figure 4_1 C). To determine whether these fusion enzymes can convert a single genomic cytidine to thymidine *in vivo*, we integrated a single-copy GFP reporter into the *E. coli* bacterial genome by recombineering (*13*) (Figure 4_1 B) in which an impaired start codon (ACG) was designed upstream of the ZF binding sequence and the GFP coding sequence. Correction of the genomic ACG to ATG by targeted deamination should result in translatable GFP transcripts and GFP-positive cells, allowing successful targeted deaminase activity to be measured by flow cytometry.

Among the four chimeric deaminases we tested, one (ZF-AID) led to robust GFP expression in the reporter population; 10 hours after ZF-AID induction 0.1% of the cell were GFP+ (Figure 4_1 C). This frequency was more than fifteen-fold higher than when ZF or AID was expressed alone (t-test, two-tailed, $P_{(ZF\text{-}AID,\ ZF)} = 0.0015$, $P_{(ZF\text{-}AID,\ AID)} = 0.0016$; n=4) (Figure 4_1 C). We confirmed with sequencing that the broken start codon ACG was permanently changed to ATG in the *gfp* gene of 20/20 randomly chosen GFP+ colonies. We conclude that AID can effectively introduce C→T mutations at a sequence specified by a fused DNA-binding module, and so we used AID as the deaminase module in all subsequent experiments.

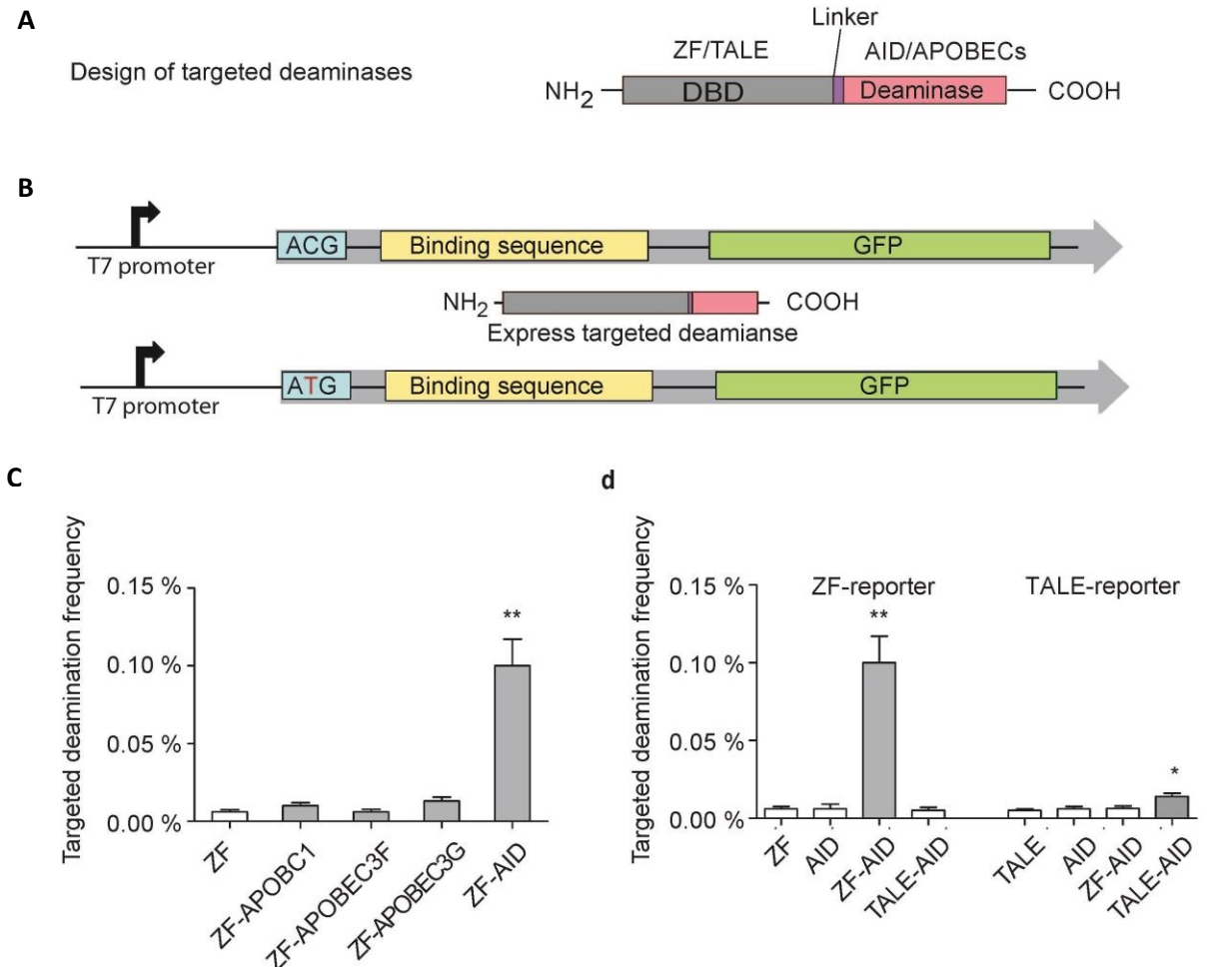


Figure 4_1 Design and targeted deaminase activity of chimeric deaminases in *E.coli*.
(A) Schematic representation of the design of targeted deaminases. The DNA binding domain (DBD), either ZF or TALE, was fused to N-terminus of the deaminase with a certain linker.
(B) Experimental overview: we integrated a GFP cassette (top) consisting of a broken start codon ACG, DNA binding sequence, and the GFP coding sequence into the bacterial genome. We subsequently transformed targeted deaminases (middle) in pTrc-kan plasmid (Supplementary Method1) into the strain and induced protein expression. Targeted deamination of the C in the broken start codon leads to a ACG→ATG transition (bottom), rescuing GFP translation which is quantifiable via flow cytometry.
(C) ZF-deaminases were tested for targeted deaminase activity by measuring GFP rescue. ZF, ZF-APOBECs (ZF-APOBEC1, ZF-APOBEC3F, ZF-APOBEC3G) or ZF-AID indicate cells transformed with plasmids that express ZF, ZF-APOBECs or ZF-AID respectively. All error bars indicate s.d. (All t-tests compare ZF-deaminases against the ZF control. Pvalue < 0.05 *, Pvalue < 0.01 **, Pvalue < 0.001 ***, n=4).
(D) GFP rescue by ZF-AID and TALE-AID in the ZF-reporter and TALE-reporter strains.(All t-tests compare the fusion deaminases against the AID control. Pvalue < 0.05 *, Pvalue < 0.01 **, Pvalue < 0.001 ***, n=4).

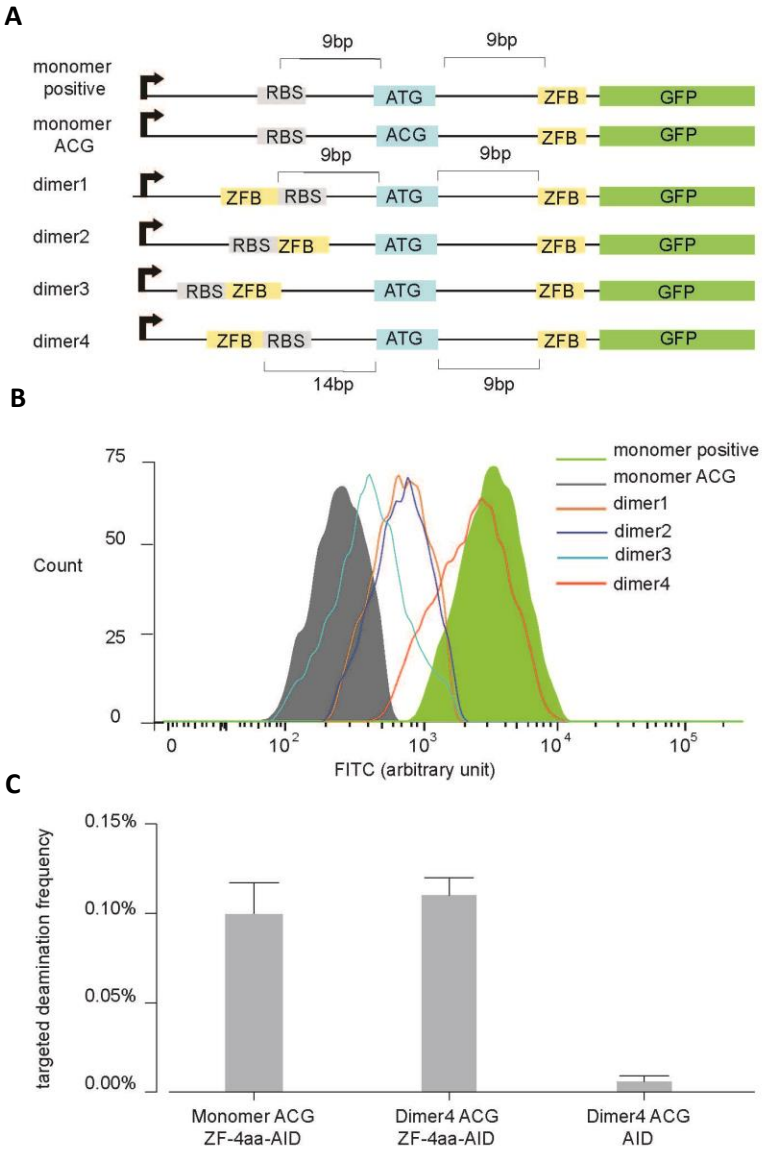


Figure 4_2 Test of targeted deaminase frequency on the reporter with two ZF binding sites.
(A) Schematic representation of the modified GFP reporters with two ZF binding sites. In the monomer reporter, a ZF binding site (ZFB) lies 9bp downstream of the start codon (in blue). In the dimer reporter constructs, an additional ZFB lies either 9bp (dimer1 and dimer3), 6bp (dimer2), or 14bp (dimer4) upstream of the start codon. Arrows indicate promoter, RBS indicate position of ribosome binding site.
(B) Overlap histogram of GFP expression level from the different reporters. Dimer1, 2 and 3 exhibited significant overlaps with the negative control (uninduced monomer ACG reporter), suggesting that the alterations to the length or sequence between the RBS and start codon compromised the translation of GFP. In contrast, the dimer4 reporter showed distinct GFP fluorescence, so we chose it for the following test.
(C)Targeted deamination frequency on dimer and monomer reporters. ZF-4aa-AID expression led to similar GFP rescue frequency in both the dimer4 ACG and monomer ACG GFP reporter systems. Conversely, AID expression alone did not result in any detectable GFP rescue signal, indicating that the ZF-4aa-ZFP monomer was able to specifically target the genomic site. Targeted deamination frequency was quantified via percentage of GFP-expressing cells in the population.

We then tested whether if we could program the DNA-binding specificity of targeted deaminases by changing the DNA-binding modules. To this end, we constructed a TALE-AID fusion, using a TALE reported to recognize the 14bp sequence 5'-TCACGATTCTTCCC-3' (*14*) and built a corresponding reporter strain with the TALE binding site downstream of the GFP broken start codon (ACG) (Figure 4_1C). Induction of TALE- AID for 10 hours led to successful GFP expression in 0.02% of the reporter population (Figure 4_1 D), lower than in the ZF-AID experiment but still significantly higher than with TALE or AID expression alone (t-test, two-tailed, $P_{(TALE\text{-}AID,\ TALE)}$ =0.0069, $P_{(TALE\text{-}AID,\ AID)}$ =0.0186; n=4) (Figure 4_1 D). In addition, both TALE-AID and ZF-AID caused minimal GFP expression when their recognition sites were absent (Figure 4_1 D). Thus, both ZF and TALE DNA-binding modules can direct deaminase activity to a sequence-defined locus in the genome. Additionally, we tested whether an additional binding sequence might enhance targeting frequency by inserting a second binding site upstream of the targeted cytidine. However, we did not observe an increase (Figure 4_2).

Optimization of genome editing efficiency of targeted deaminase in bacteria

Our initial tests demonstrated the feasibility of using targeted deaminases for genome editing, but the editing efficiency was low. We reasoned that native uracil repair pathway might prevent targeted cytidine deamination from leading to a C:G→T:A transition. Therefore, we knocked out *mutS* and *ung*, two genes known to be involved in AID-initiated deamination repair. GFP rescue frequency by ZF-AID increased to 0.5% (5-fold) in the Δ*mutS* knockout, and to 3.5% (35-fold) in the Δ*mutS* Δ*ung* double knockout (Figure 4_3). Similarly, GFP rescue by TALE-AID induction was increased to 0.1% (7-fold increase) in the Δ*mutS* Δ*ung* knockout (Figure 4_3). We confirmed the GFP fluorescence signal by microscopy (Figure 4_3) and

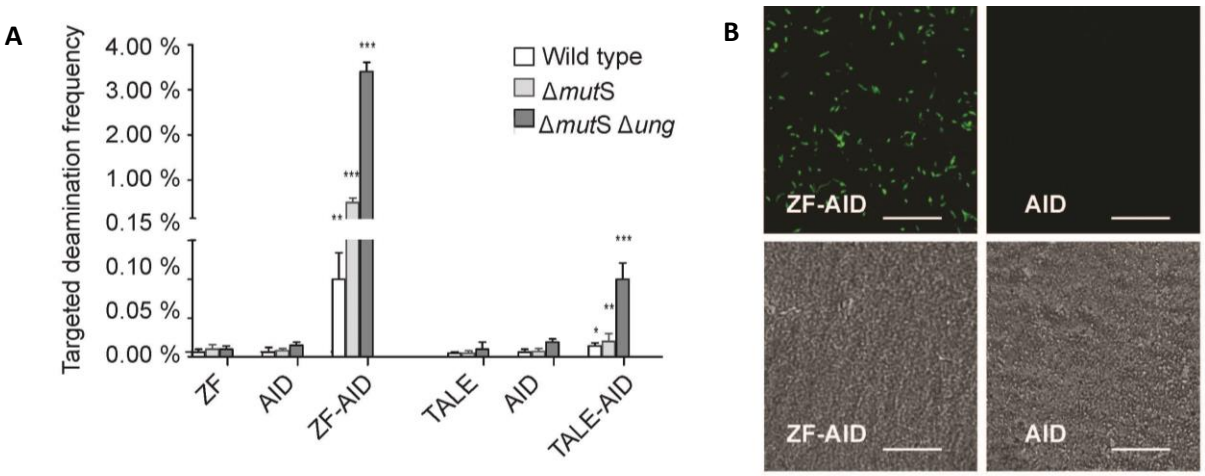


Figure 4_3. Modified genetic banckground spurs the genome editing efficiency
(A) GFP rescue by ZF-AIDs and TALE-AID in (wild type), (Δung), and (ΔmutS Δung) strains. All error bars indicate s.d.. (All t-tests compare the fusion deaminases against the AID control. Pvalue < 0.05 *, Pvalue < 0.01 **, Pvalue < 0.001 ***, n=4).
(B) E.coli (ΔmutS Δung) cells imaged under fluorescence(upper) and phase contrast(lower) after expression of ZF-AID or AID for 10 hours. Top, Scale bar: 20μm.

confirmed the C:G→T:A transitions by sequencing the *gfp* gene of 20 randomly chosen GFP+ colonies from both the ZF-AID- and TALE-AID-induced population. We conclude that suppression of uracil repair led to an increase in the resulting rate of C:G→T:A transitions caused by targeted deaminase. Hence, all subsequent experiments with *E.coli* were done in the Δ*mutS* Δ*ung* background.

As genome engineering requires both efficiency and specificity, we then set out to increase editing efficiency via structural optimization of the fusion enzymes. First, we compared the ZF-AID described above with three alternatives carrying modified linkers previously used in zinc finger fusion proteins (*15, 16*) (an alternative four amino-acid linker, 4aa2, to examine the effect of linker sequence, and 8aa and 11aa linkers to examine the importance of linker length). (Figure 4_4). While expression of all four ZF-AIDs led to robust GFP rescue, improvements were observed with ZF-8-aa-AID achieving 7.5% GFP+ frequency after 10 hours (Figure 4_5), and 13% after 30 hours of induction. While linker length is clearly an important design consideration, interestingly, rescue efficiencies by ZF-4-aa-AID and ZF-4-aa2-AID were also slightly different (t-test, two tailed, p=0.0032, n=4), suggesting that besides linker length (4 amino acids in both), linker sequence also influences performance of the overall construct.

The initial test with TALE –AID (hereafter referred to as TALE-C1-AID) did not rescue GFP with as a high frequency as the ZF-AIDs. Given the critical importance of linker length observed with our ZF-AID fusion constructs (Figure 4_4), we proceeded to examine whether truncation of some or all of the 178aa region in the C-terminus of the TALE protein could enable a higher GFP rescue frequency. To this end, we tested the activity of TALE-AIDs carrying various truncated TALE C-termini (referred as TALE-C2-AID, TALE-C3-AID, TALE-C4-AID, and TALE-C5-AID (Figure 4_6)). Truncations were chosen at loop regions predicted

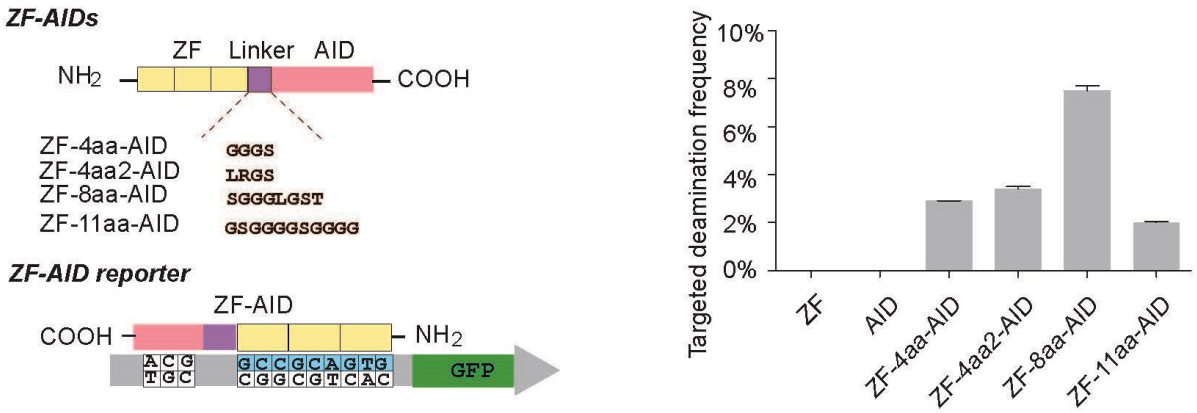


Figure 4_4. Optimization of ZF-deamination s frequency in E.coli.
Schematic representation of ZF-AIDs variants tested for targeted deaminase activity (upper) and the reporter (lower) with the ZF-recognition sequence in blue. b, GFP rescue by expression of the four ZF-AIDs variants and ZF or AID domains alone. All error bars indicate s.d..

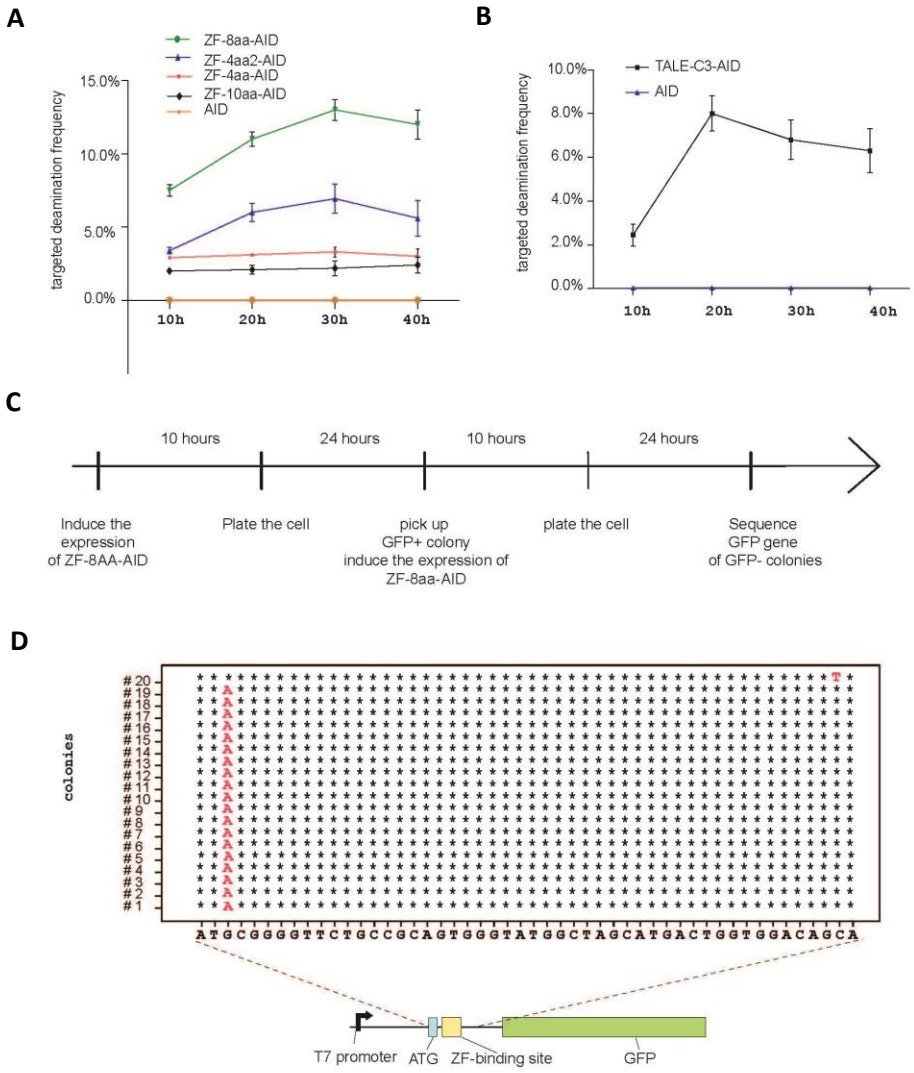


Figure 4_5. Secondary mutations led to the decline of GFP rescue efficiency

(A) Targeted deamination frequency peaked following 30 hours of ZF-AID induction and dropped after that. The targeted deaminase frequencies were measured by flow cytometry analysis of GFP expression. Bacterial culture was diluted 1:100 every 10 h to maintain continuous cell proliferation.

(B) Targeted deamination frequency as measured by GFP+ cell fraction peaked following 20 hours of TALE-AID induction and dropped after that. Bacterial culture was diluted 1:100 every 10hrs to maintain continuous cell proliferation.

(C) Time line depicting the experiment design to capture secondary mutations.

(D) Sanger DNA sequencing revealed that prolonged ZF-AID induction led to secondary mutations that abolished the expression of GFP. 1kb of the *gfp* gene was sequenced over 20 GFP- colonies; only the mutated part is shown in the table. The original sequence is listed below and the schematic graph of the GFP cassette shows the corresponding positions of this sequence. "*" indicates positions where the sequence is identical with the wild type *gfp*. Red letters indicate the mutated bases.

using protein secondary structure by the software LASERGENE. Changes in TALE C-terminus length could affect targeted deaminase activity at a particular target locus either by affecting intrinsic protein activity or simply by making the protein optimal for a different length of DNA between the DNA binding and deamination target sites. To investigate these possibilities, we also constructed five bacterial GFP reporter strains, each with a genomic *gfp* locus carrying a broken start codon 2, 5, 8, 11, or 14bps upstream of the TALE binding site (Figure 4_6). Targeted deamination frequencies were then measured by GFP rescue frequency and compared in a 5-by-5 matrix of TALE-AIDs and reporters (Figure 4_6). TALE-AID truncations showed significantly higher GFP rescue over that of TALE-C1-AID (Figure 4_6). Notably, induction of TALE-C3-AID achieved a genomic editing frequency of 2.5% on the 8bp-spacer reporter after 10 hours of induction (Figure 4_6), and peaked at 8% following 20 hours of induction (Figure 4_5 B). Interestingly, TALE-C3-AID outperformed all other constructs regardless of the spacer length of the reporter, suggesting that this chimeric protein has an intrinsically optimal structure out of the TALE-AIDs tested. Taking together the optimization results for ZF-AIDs and TALE-AIDs, we suggest important design considerations for engineering of efficient targeted deaminases.

Test of the specify of targeted deaminase in bacteria

Having investigated and improved deaminase targeting frequency, we next characterized targeting specificity using the following three methods: 1) investigating the effect of point-mutations in the recognized DNA sequence on editing frequency 2) sequencing the GFP locus in many cells after expression of targeted deaminase; and 3) whole-genome sequencing of three GFP+ clones.

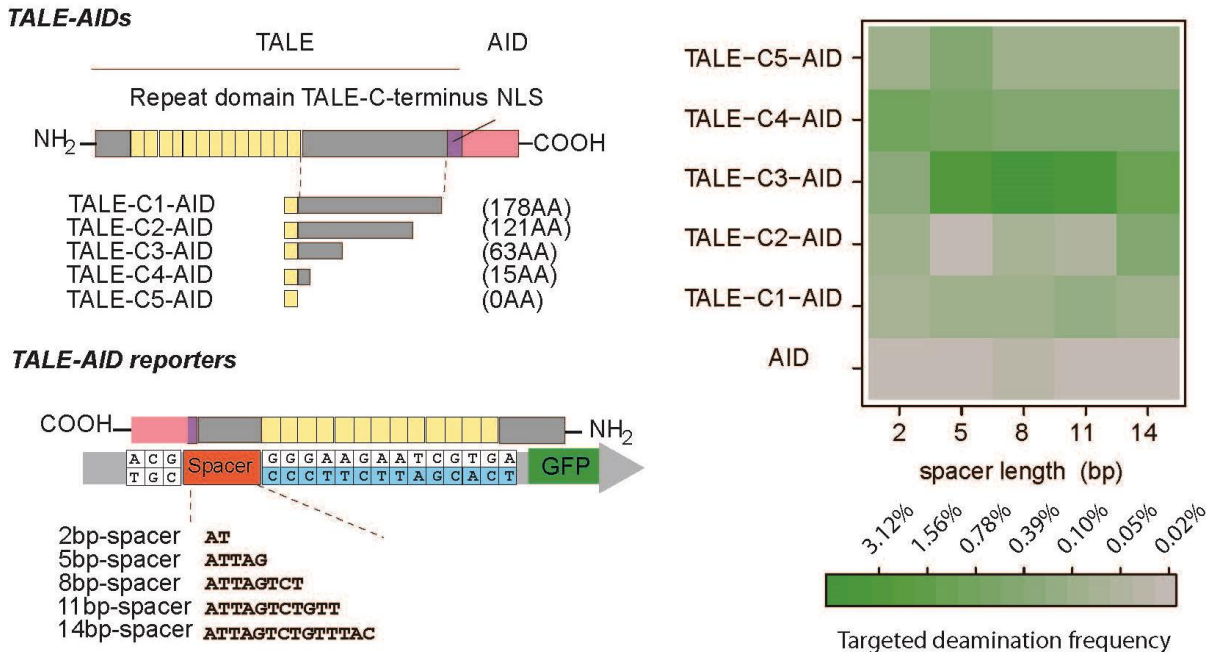


Figure 4_6. Optimization of TALE-deamination s frequency in E.coli.
Schematic representation of TALE-AIDs and the reporters tested for targeted deaminase activity. Five TALE-AIDs (upper) with different TALE C-terminus truncations (C1 to C5) were constructed, with the remaining C-terminus lengths shown in parentheses. Full TALE-AID protein sequences can be found in Supplementary Sequence 2. Five reporters were constructed (lower) with different spacer lengths (2bp, 5bp, 8bp, 11bp) between the broken start codon and TALE DNA binding motif. The TALE binding site on the GFP reporter is shown in blue; the TALE N-terminus segment specifies the 5′ thymine base of the binding site. d. All five TALE-AIDs were tested for targeted deaminase activity on all five reporters . Green and grey encode high and low GFP rescue, respectively.

We tested DNA sequence specificity of targeted deaminases by measuring GFP rescue using reporters with point-mutated ZF/TALE recognition sequences. We first altered three individual nucleotides within the nine-nucleotide ZF recognition sequence. Divergence from the intended recognition sequence by a nucleotide led to 4-8 fold decrease in ZF–8aa-AID efficiency (Figure 4_7 A), indicating that ZFP-8aa-AID is highly specific to the ZF-addressed locus. We next investigated the specificity of TALE-AID by individually mutating each nucleotide in the TALE recognition site to the second most preferred base for that position (Figure 4_7 B). Interestingly, TALE-C3-AID, which was designed to recognize a 14bp sequence, showed strong sequence specificity only for the first 8bp proximal to the target site (5' TTCTTCCC 3' in the TALE recognition site). Thus, both ZF- and TALE-AID demonstrate sequence specificity. However, for reasons that remain to be investigated, sequence alterations at more distal positions in the TALE binding site led to variable targeting frequency (Figure 4_7 B).

To detect possible off-target mutations close to the intended deaminase target site, we sorted 10,000 GFP+ and 10,000 GFP- cells after 30 hours of ZF-8aa-AID induction, and randomly isolated 200 individual colonies from each population. We Sanger sequenced 1kb surrounding the *gfp* deaminase target site and, as a control, the constitutively expressed *gapA* gene, which lies 1.9Mbp away from *gfp*. In the GFP+ population, all colonies harbored the intended C→T transition in the *gfp* start codon. 5.5% of these colonies contained an additional C→T mutation upstream or within the *gfp* gene (Figure 4_7 C). In the GFP- population, the only mutation detected over 200 colonies was a single G→A transition 1bp away from the intended target site (ACG→ACA), present in 2% of the population (Figure 4_7 C). No mutations were found in *gapA* in any colonies from the two populations. AID has been documented to have sequence preference, targeting cytidines in its "hotspot" (WRC, W=A/T, R=A/G) more

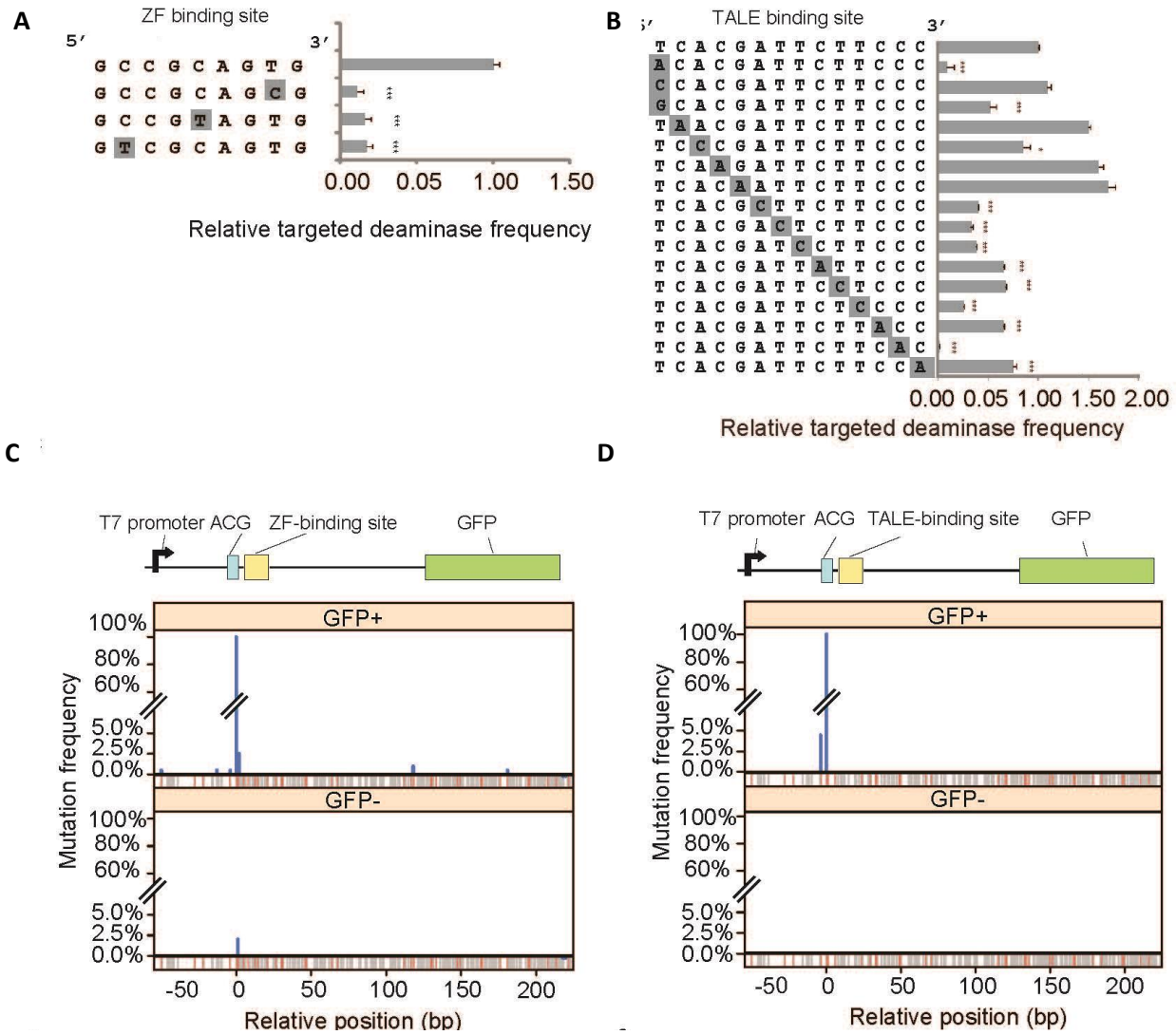


Figure 4_7. Test of the specificity of AID fusions.

(A) Test of ZF-8aa-AID sequence specificity using a GFP reporter with point-mutated ZF binding sequences. t-tests compare each mutated site against the unmodified site (top). Pvalue < 0.05 *, Pvalue < 0.01 **, Pvalue < 0.001 ***, n=4. All error bars indicate s.d..

(B) Test of TALE-C1-AID sequence specificity using a GFP reporter with point-mutated TALE binding sites. t-tests compare each mutated site against the unmodified site (top). Pvalue < 0.05 *, Pvalue < 0.01 **, Pvalue < 0.001 ***, n=4. All error bars indicate s.d.. Note that we altered the first nucleotide, a TALE-N terminus-specified thymine, to other three nucleotides individually, while we changed other nucleotides in the TALE recognition domain to the nucleotide mostly likely to be recognized5.

(C) Mutation location and spectrum in the GFP gene of GFP+ and GFP- cells collected after ZF-8aa-AID induction. A schematic structure of the GFP gene is shown above the mutation frequency along the gene's length among 200 Sanger sequenced colonies of each cell population. Gray lines indicate positions of C/G nucleotides; red lines indicate occurrences of the AID preferred motif (WRC).

(D) Mutation spectrum on the GFP gene of GFP+ and GFP- cells collected after TALE-C1-AID induction

frequently than other sites (*17*). It has also been reported that AID exhibits processive behavior, tending to deaminate a stretch of cytidines on the same DNA strand (*18*). Interestingly, only 0.7% of non-target WRC sites in the *gfp* locus were altered in the GFP+ population (Figure 4_7 C), and these mutations were not clustered at any single read. This suggests that ZF targeting overrides the native sequence preference as well as processive behavior of the AID enzyme. We next repeated our assay using TALE-C3-AID. In the GFP+ population, besides the intended C→T mutation, an additional C→T mutation 4bp upstream of the intended site was found in 9/200 colonies (4.5%) (Figure 4_7 D). No other off-target mutations were detected in the GFP coding sequence or in the GFP- cells. Given that no mutations were found in the distant *gap*A sequence and that off-target mutations were enriched in the GFP+ versus GFP- cells, we speculate that the observed off-target mutations might be caused by residual processivity of AID, the flexibility of the linker and one-dimensional sliding of the DNA binding protein along the chromosome (*19*). Taking ZF- and TALE- deaminase together, among cells that were correctly engineered (GFP+), approximately 95% had no off-target modification in the *gfp* locus. The only observed off-target modification frequency in GFP- cells was 1bp away from the intended site, at a frequency of 2%. These results indicate that the deaminase activity was tightly constrained by the ZF and TALE DNA-binding modules in the fusion proteins.

To more completely assess the global off-target activities of the fusion proteins, we sequenced with ~50X coverage the genomes of three GFP+ colonies edited by ZF-8aa-AID, and three colonies edited by TALE-C1-AID, and compared them each to control GFP- colonies in which the expression of deaminases had not been induced. We did not find a significant elevation of nucleotide substitutions in the edited clones relative to the uninduced control

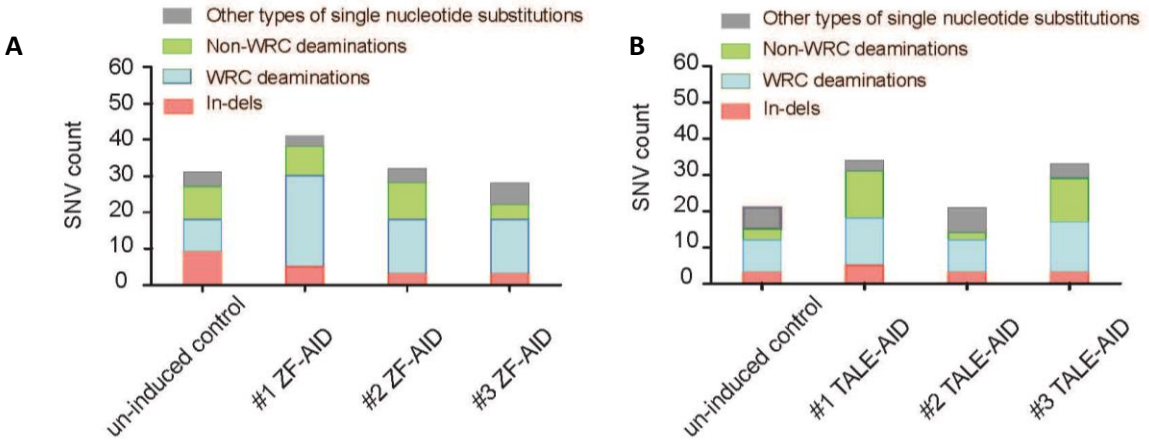


Figure 4_8. Unbiased test of specificity of AID fusion via whole genome sequencing of bacteria
(A) Whole-genome SNV profiles of strains with/without ZF-AID induction. SNVs that may stem from cytosine deamination (C/G→T/A) are in either green (if C was in the AID-preferred WRC motif) or blue (all other Cs) bars.
(B) Whole-genome SNVs profiles of strains with/without TALE-AID induction. Color schematic is the same as 3e.

(Wilcoxon test, P_{value}=0.25) (Figure 4_8). C:G→T:A transitions, likely due to cytidine deamination, were elevated relative to other mutations in both experimental and control groups, as expected from Δ*ung* Δ*mutS* knock-outs. ZF-AIDs triggered a subtle enrichment in WRC deamination (Figure 4_8 A), and two of the TALE-AID colonies incurred slightly increased WRC and non-WRC deaminations in their genomes (Figure 4_8 B). Nevertheless, the overall C:G→T:A rate under ZF-AID/TALE-AID expression remained close to that of the uninduced strains (uninduced: 15±4 deaminations, induced: 23±8 deaminations; Figure 4_8), suggesting that global off-target deamination occurred to a small extent, if at all. No enrichment of in-dels and no structural rearrangements were detected in any of the, suggesting that ZF/TALE-AID in the Δ*ung* Δ*mutS* background did not trigger DSBs or subsequent NHEJ. Of note, there are other four sites in the genome containing the exact ZF binding sequence with a WRC motif 9bp upstream. All of these cytidines remained un-mutated, but this is statistically consistent with expectations given the number of sites and the measured target site editing frequency (Methods). However, we cannot exclude the possibility that other factors, like local transcriptional activity (Figure 4_9) or genomic position effects, might differentially affect the editing efficiency at these five sites. Taken together, whole-genome sequencing showed that ZF/TALE-AID is specific to the target site and does not cause hypermutation or genomic structural changes.

Test of the activity and toxicity of targeted deaminases in human cells

Given the intense interest in achieving facile genomic editing for human studies, we tested if our targeted deaminase system would function in human cells. We constructed a reporter for human cells in which an EF1α promoter drives expression of a broken-start-codon

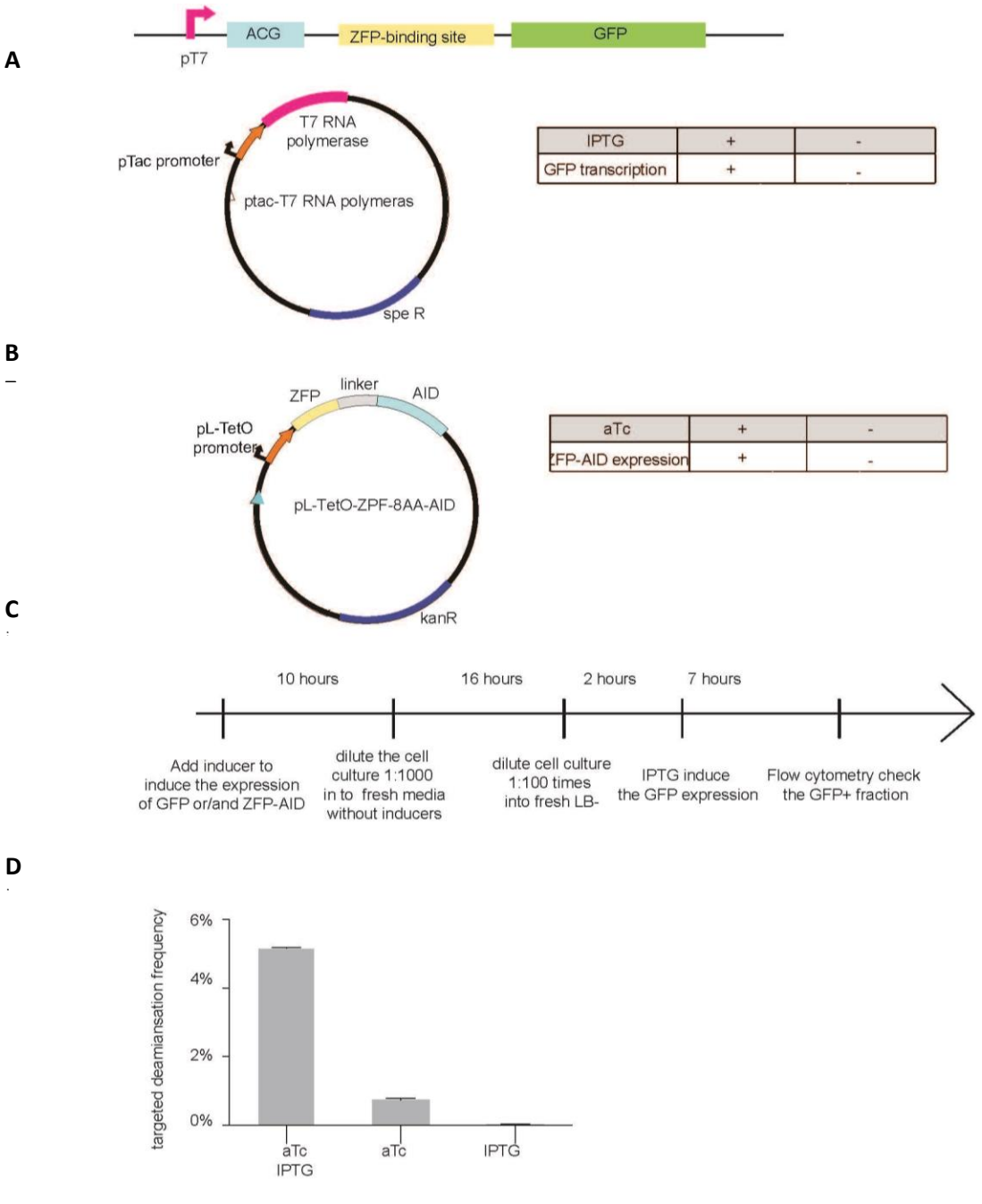
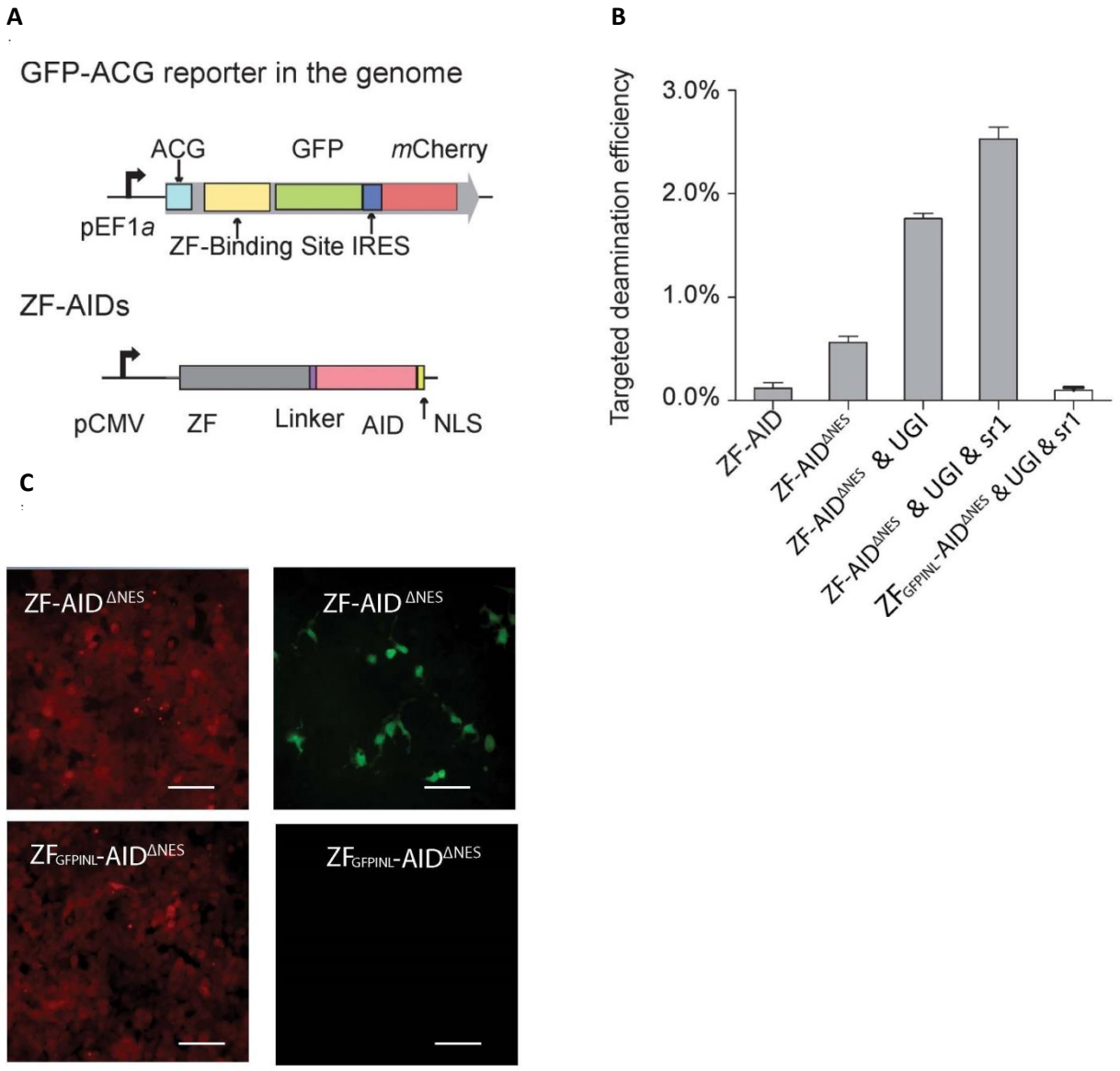


Figure 4_9 Active transcription enhances targeted deamination.
(A) Schematic representation of the transcription control of the GFP reporter. This GFP was transcribed by T7 RNA polymerase which is transcribed by an IPTG inducible promoter pTac.
(B) Schematic representation of the transcription control of ZF-AID. ZF-AID was transcribed from the pL-TetO promoter which was modulated by the TetR protein (constitutively expressed) and the inducer aTc.
(C) Time line depicting the experiment design.
(D) Targeted deamination frequency with/without GFP transcription. The bacterial culture was induced with IPTG, aTc and IPTG&aTC for 10 hours, and then diluted 1000-fold into fresh media without any inducer overnight. Cell culture was diluted again 100-fold into fresh media with IPTG to check for the expression of GFP. Targeted deamination frequency was quantified via percentage of GFP-positive cells in the population.

(ACG) GFP attached to an IRES-mCherry selection marker. We stably inserted this construct into HEK293FT cell lines by lentiviral transduction and established a monoclonal cell line by FACS sorting (Figure 4_10 A). The optimized ZF-AID construct (ZF-8aa-AID) was then delivered into the reporter cell line via transfection (Figure 4_10 A). As expected, 48 hours of expression of ZF-8aa-AID led to GFP expression in 0.12% of transfected cells. We next constructed ZF-AID$^{\Delta NES}$ by truncating the 15aa from the C-terminus of AID, which contains a nuclear export signal (*20*) and regions that interact with mismatch repair proteins (*21*). This is expected to: 1. Correctly localize ZF-AID to the nucleus; 2. Contribute to editing success via decoupling AID from mismatch repair; 3. Minimize toxicity caused by repair-associated DSBs[24]. As expected, expression of ZF-AID$^{\Delta NES}$ significantly increased GFP rescue versus that of full-length ZF-AID (Figure 4_10 B) (0.56%, t-test, two-tailed, n=4, P_{value}=0.0013). As in *E.coli*, we then tested the effect of suppressing uracil repair by using the UNG inhibitor UGI (*22*) and knocking down *MSH2* (the human homolog of bacteria *mutS*) with shRNA. UNG and MSH6 suppression together increased ZF-AID$^{\Delta NES}$–mediated GFP rescue efficiency to 2.5% (Figure 4_10 B). Expression of ZF_{GFPINL}-AID$^{\Delta NES}$, a fusion protein whose zinc finger domain targets a site 265bp from the GFP start codon, resulted in minimal GFP rescue (Figure 4_10 A, C), suggesting both that genome editing by ZF-AID$^{\Delta NES}$ is sequence-specific, and that the ZF-binding site and the targeted cytidine must be close for measurable activity. Successful C:G→T:A targeting of the broken start codon was confirmed by Sanger sequencing of the *GFP* locus in 8/8 stable GFP+ colonies. Therefore, engineered deaminases are capable of efficient sequence-specific genome editing in HEK293 cells. We next sought to characterize the toxicity of targeted deaminase in human cells. This is important as AID has been implicated in contributing to translocation-associated lymphomas



.

Figure 4_10 Targeted deamination of ZF-AID in human cells

(A) Schematic representation of the ACG-GFP reporter system in HEK239FT cells (upper) and the ZF-AID (lower) tested for targeting deaminase activity. IRES, Internal ribosome entry site; NLS, nuclear localization signal.

(B) Targeted deamination activity of ZF-AIDs. ACG-GFP reporter cells were transfected with the constructs labeled on the X-axis. Targeted deamination frequency was estimated as the proportion of GFP-rescued cells 48h after transfection. ZF-AIDΔNES is identical to ZF-AID except with a deleted AID nuclear export signal (NES); UGI, inhibitor of UNG; sr1, shRNA-MSH2.

(C) ACG-GFP reporter cells imaged under fluorescence (mCherry (left)/GFP (right)) 48hr after transfection with ZF-AID ΔNES /UGI/sr1 or ZF GFPINL-AIDΔNES /UGI/sr1 plasmids. Scale bar = 200μm

(*23*) by recruiting repair machineries to create staggered DSBs (*21*). To test whether ZF-AID$^{\Delta NES}$ can be safely used as a genome editing tool without incurring DSBs, we generated a HEK293FT reporter cell line carrying a non-functional frame-shifted GFP, which could be rescued by DSB-induced HR with exogenous donor DNA (*5*, *6*). DSB frequency caused by a particular cell treatment can therefore be estimated by the frequency of GFP+ cells generated by the treatment. The GFP-In reporter also carried recognition sites for two known DSB-creating proteins, I-*Sce*I and ZF$_{GFPIN}$Ns (ZF$_{GFPINL}$N & ZF$_{GFPINR}$N), for use as positive controls (*6*). While expression of I-*Sce*I and ZF$_{GFPIN}$Ns generated 1.01% and 0.43% GFP+ cells respectively, the result for ZF$_{GFPIN}$-AID$^{\Delta NES}$s was just 0.03%, which was reduced to 0.01% if the UNG inhibitor UGI was co-transfected (Figure 4_11 B), consistent with previous observation (*21*, *24*). Thus, combined targeted deaminase and UGI treatment created 40-fold fewer DSBs at the target locus than the zinc finger nuclease treatment, close to the level of a negative control where only the DNA donor was delivered. The extent to which ZF-AID$^{\Delta NES}$ generated DSBs (0.01%) was very low compared to its C:G→T:A editing activity (2.1%) (Figure 4_10 B). Furthermore, we observed higher cell survival in the ZF$_{GFPIN}$-AID$^{\Delta NES}$s/UGI-expressing population (66%) than in the ZF$_{GFPIN}$Ns-expressing population (41%, Figure 4_11 C), suggesting that targeted deaminases are less toxic than ZFNs. Thus, expression of chimeric AID$^{\Delta NES}$s with UGI enables efficient genomic editing in human cells without generating DSBs and with low cytotoxicity.

Discussion

Our study demonstrates that fusing cytidine deaminases with DNA binding modules enables site-specific deamination of genomic loci in both prokaryotic and eukaryotic cells. We designed and optimized the structure of targeted deaminases to effectively convert a specific C:G

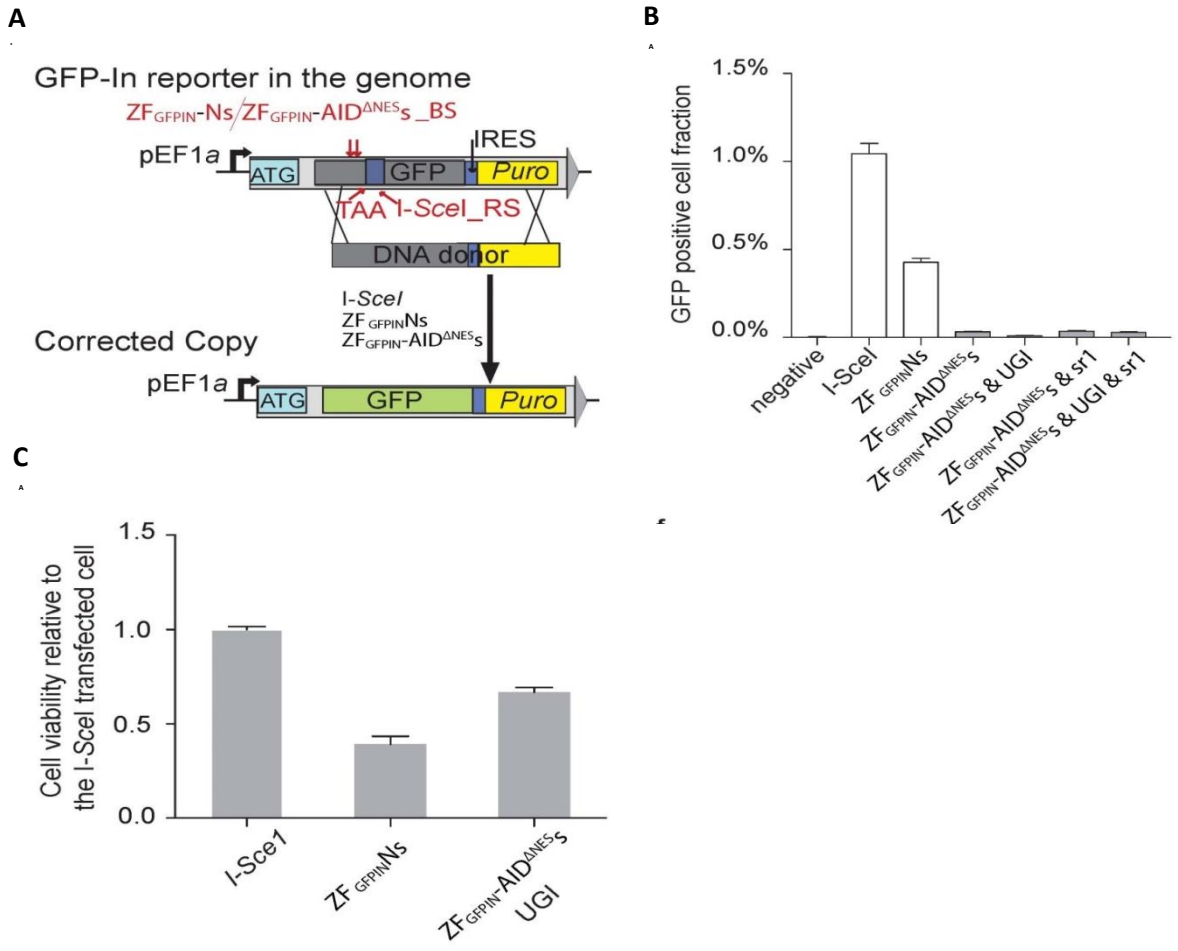


Figure 4_11 Targeted deamination of ZF-AID in human cells.
(A)Schematic design of DSBs assay. The genomically integrated GFP-In reporter includes a 35bp frame-shift insertion bearing a stop codon and I-SceI recognition site (I-SceI_RS). Of note, ZFGFPINNs and ZFGFPIN-AIDΔNESs binding sites (ZF GFPINNs/ZFGFPIN-AIDΔNESs_BS) were identical and located 82bp upstream of the insertion. We transfected the cells with a DNA donor carrying the wild-type GFP sequence along with I-Sce1/ ZFGFPINNs / ZFGFPIN-AIDΔNESs expression plasmids and assessed the DSB-generating rate by measuring HR frequency as determined by GFP rescue of the cells.
(B) GFP rescue results determined by flow cytometry. Negative control was transfected with the DNA donor only.
(C) Cytotoxicity assay for ZFGFPIN-AID/UGI relative to I-Sce1. Detailed methods are in Methods. A value of <1 shows decreased cell survival as compared to I-SceI, and demonstrates a toxic effect.

base pair to T:A in the *E.coli* genome, achieving 13% editing frequency, 95% local targeting accuracy and a low rate of genome-wide off-target mutations. We then applied the optimized chimeric deaminases to a human cell line and found that these novel enzymes could create site-specific single-nucleotide transitions in as many as 2.5% of transfected cells. Targeting activity rarely creates DSBs and led to increased cell survivability relative to the zinc finger nuclease editing method that is currently broadly employed. These results set the stage for the future engineering of additional desirable functionalities onto the engineered targeted deaminases, including targeted adenosine deaminases (*25*), hyperactive and processive targeted deaminases. Such a 'molecular toolkit' for targeted genome editing will find numerous biotechnological and therapeutic applications (*26, 27*).

Further investigation is needed to address remaining questions: for instance, it is currently unknown whether AID acts as a monomer , or a dimer (*12*). While we have found that binding of two ZF-AIDs flanking the target site does not increase modification rates compared to binding of a single ZF-AID (Figure 4_2), it is possible that a single ZF-AID recruits a second ZF-AID to the target site through free dimerization of AIDs. To the extent that dimers are required for functionality, targeting of deaminases might be improved by engineering obligate heterodimers as has been done for ZFNs (*28*) . Additionally, the dependence of ZF/TALE-AID activity on the transcriptional state of the targeted gene requires further characterization. If transcription at the desired locus is essential for targeted deamination, combining programmable transcriptional activators with targeted deaminases may allow effective editing at transcriptionally silent sites.

While targeted deaminases enable only the generation of single base transitions in a genome compared to the broader capabilities supported by other methods, this is sufficient for

important applications such as creation of nonsense and missense mutations at defined genomic loci, and this limitation is offset by several advantages, notably simplicity, self-contentedness, low toxicity, and high potential for multiplexing. Unlike targeted nuclease-based methods, targeted deaminases do not depend on complex DNA repair pathways operating in the cells or on the differential induction of competing repair pathways such as HR vs. NHEJ. This not only makes them simple to use but also removes a significant barrier to their portability to other cell types and organisms, as seen by the ease with which our system was moved from E. coli to human. Indeed, by contrast, oligo-mediated genome engineering has proved difficult to port from E. coli to human even though the Lambda-Red co-factors needed for the former are well defined and few in number (*13*). Moreover, unlike targeted nucleases, targeted deaminases do not depend on DSB repair, and so should neither induce error-prone NHEJ nor cause off-target DSBs that result in cytotoxicity. This suggests that, compared with targeted nucleases, many targeted deaminases could be expressed in a cell to effect multiplexed changes without toxicity, and that deaminases could be targeted to edit abundant sequences in the cell such as common regulatory sequences or even repetitive elements. Finally, the lack of need for exogenous donor DNA potentially makes targeted deaminases attractive for use in multicellular organisms. We envision experiments in which targeted deaminases could be introduced into intact animals via safe-harbor knock-ins or viral vectors, and turned on at specific times in specific cell types to specifically mutate target genes seamlessly – requiring neither delivery of high copy number donor DNA for targeted nucleases, which is challenging, nor emplacement of complex DNA constructions including recombinase sites (such as loxP sites) that introduce unwanted DNA sequences and leave scars on excision. Because of these features, we expect that targeted

deaminases will prove to be effective tools for genome engineering, and we plan to explore these applications in future experiments.

Materials and Methods

Construction of fusion proteins

To construct ZFP-AID fusion proteins, we first PCR amplified ZFP from pUC57-ZFP12 and AID from pTrc99A-AID31 and fused these two parts with various linkers using overlap PCR. The fusion constructs were cloned into a pTrc-Kan plasmid. We fused AID with TALE by cloning AID into pLenti-EF1a-TALE(0.5 NI)-WPRE15 plasmid and then cloned TALE-AID fusions into the pTrc-Kan plasmid. APOBEC1, 3F, and 3G genes were synthesized (Genescript) and cloned into the pTrc-ZFP-Kan plasmid. To generate pCMV-ZF-AID constructs, we amplified ZF-AID cassette from pTrc-ZF-AID and cloned that into pCMV-hygo15 plasmid.

1. Restriction enzymes and Rapid ligase were purchased from New England Biolabs and used according to the manufacturer's instructions. PCRs were conducted by Kapa HiFi PCR 2X master mixture (Kapa Biosystems). The primers and oligos were obtained from IDT and gene synthesis was provided by Genescript. All of the primer, construct and backbone sequences are listed in Supplementary Sequences.

2. Construction of pTrc-Kan as the backbone vector

3. We first constructed a common inducible expression vector by combining the elements from pZE-21 and pROEX-HTa vectors. In brief, the fragment containing lacI gene and pTrc promoter of pROEX-HTa was amplified by PCR using lacI-XhoI and pTrc-HindIII primers. This fragment was digested with Xhol and HindIII and ligated into a similarly digested

pZE-21 to make pTrc-Kan. A NheI restriction site was also imbedded downstream of the pTrc promoter for future cloning.

4. Construction of pTrc-ZF and pTrc-AID

The ZF gene [10] was amplified from pUC58-ZFP by PCR using ZF-F and ZF-R-HindIII primers. ZF fragments were digested with HindIII and NheI and ligated into the pTrc-Kan backbone plasmid, which was similarly digested. The AID gene was amplified from pTrc99A-AID, a gift from Meng Wang[11], using primers AID-F-NheI and AID-R. AID fragments were digested with HindIII and NheI and ligated into the pTrc-Kan backbone plasmid digested with the same enzyme.

5. Construction of pTrc-ZF-AIDs

The ZF was appended to the N-terminus of AID using amino acid linkers of various sizes and composition. The ZFP gene was amplified from pZFPerb2 and the linker sequence was created by PCR using ZFP-F and ZFP-R (4aa, 4aa2, 8aa, 11aa) primers individually. In parallel, the AID gene was amplified from ptrc99A-AID and the linker sequence was created by PCR using AID-F (4aa, 4aa2, 8aa, 11aa) and AID-R primers. ZF and AID with corresponding linkers were fused by overlap extension PCR using ZF-F and AID-R primers. Each construct was digested with NheI and HindIII and ligated into the four similarly digested pTrc-Kan backbone plasmids.

6. Construction of pTrc-ZF-APOBECs

We constructed pTrc-ZF-APOBECs with various linkers using the isothermal assembly protocol12. In brief, the APOBEC1, APOBEC3F, APOBEC3G 2K3A genes were amplified by PCR using primers APOBEC-F and APOBEC-R. pTrc-ZFP-aaAID was linearized by SalI and

HindIII digestion and the pTrc-ZF fragment was recovered by gel purification. The pTrc-ZF fragment was fused to individual APOBEC fragments by isothermal assembly.

7. Construction of pTrc-TALE-AIDs

AID gene was amplified from pTrc99A-AID plasmids, digested with NheI and BsrG1 and cloned into the pLenti-EF1a-TALE(0.5 NI)-WPRE [13], which was similarly digested. The obtained TALE-C1-AID fusion was amplified and digested with SspI and HindIII and inserted into the pTrc-Kan backbone vector to obtain the pTrc-TALE-C1-AID construct. TALE truncations were created by amplifying the TALE fragment with the appropriate TALE-F and TALE-R deletion primers. The truncated TALEs were then ligated into pTrc-TALE-AID plasmid using the SspI and NheI sites.

8. Construction of pTrc-TALE-APOBECs

APOBEC genes were each amplified by PCR using APOBEC-F-NheI and APOBEC-R-HindIII primers individually, digested with NheI and HindIII, and ligated into the pTrc-TALE-AID plasmid that was similarly digested.

9. Construction of pL-tetO-ZF-AIDs

The pL-tetO promoter was amplified from the pZE-21G plasmid by PCR using primers pL-tetO-5 and pL-tetO-3. This fragment was digested with NheI and XhoI and ligated into pTrc-ZF-AID plasmid that was similarly digested.

10. Construction of pCMV-ZF-AIDs

pCMV-ZF-AID/pCMV-ZF-AID$^{\Delta NES}$ constructs were built by amplifying ZF-AID/ZF-AID$^{\Delta NES}$ using BsiWI -ZF and BsrGI-AID/BsrGI-ΔAID primers respectively. The PCR products

were digested with BsrGI and BsiWI, and ligated into pCMV-puro backbone that was similarly digested. To generate ZF_{GFPIN}- $AID^{\Delta NES}$ expression vectors, ZF_{GFPINL} and ZF_{GFPINR} were amplified from pST1374-G223L and pST1374-G223R [14] vectors using BsiWI-ZFL/R, BamHI-ZFL/ZFR primers and cloned into pCMV-ZF-ΔAID to swap the ZF domains using BsiWI and BamHI. Subsequently, ZF_{GFPIN}- $AID^{\Delta NES}$ were amplified and cloned into pST1374 vector by NheI and ApaI restriction sits to generate ZF*-AIDs expression vectors with the same backbone and DNA binding module as ZF_{GFPIN}-Ns (ZF_{GFPINL}-Ns/ ZF_{GFPINR}-Ns)

11. Construction of pCMV-UGI

encoding gene optimized for human cell expression was synthesized and cloned into pCMV-puro using XhoI and BsiWI restriction sites. The sequences of the fusion proteins are listed in Sequence 4.1-4.5.

Construction of E.coli reporter cell lines

The GFP coding sequence was amplified from pRSET-EmGFP (Invitrogen). We modified the reporter by mutating the start codon to ACG and inserting a ZFP/TAL binding site upstream of the GFP coding sequence. To establish stable cell lines with a single copy of the GFP reporter sequence in the genome, we integrated the GFP cassette into the galK locus in the EcNR1 (MG1566 with λ-prophage::bioA/bioB) and EcNR2 (EcNR1 with mutS knocked out) strains[14]. To knock out ung, we replaced the ung gene with Zeocin resistance cassette via recombineering. In addition, all the reporter cell lines were transformed with pTac-T7polymerase to induce the expression of GFP. Subsequent modifications of the reporter were conducted using the MAGE system[14]. All sequences can be found in Sequence 4. 6 and 4.7.

1. Construction of the GFP reporter strains

First, we modified the GFP cassette on pREST-EmGFP (Invitrogen) to contain a ZFP binding site. A dsDNA fragment was synthesized with the ZFP binding site (5' GCCGCAGTG 3') 9bp downstream of a start codon, and the fragment was flanked by the NdeI and NheI restriction sites. This fragment was digested with NdeI and NheI and ligated to a similarly digested pREST-EmGFP to construct pREST-ZFP-EmGFP. Modified GFP cassette was incorporated into the galK locus in the EcNR1 and EcNR2 strains using the λ-red recombineering. In brief, the GFP cassette was amplified by PCR using 5'-galk-gfp and 3'-gfp-galk primers to create galK homology on both sides of GFP. This fragment was transformed into λ-red-induced strains, and successful insertions were selected for based on GalK negative selection [16]. Subsequently, we modified the single-copy GFP reporter through MAGE. To control the expression of GFP from the T7 promoter, we introduced the plasmid pTac-T7 RNA polymerase, in which T7 RNA polymerase is transcribed from tac promoter of the lactose operon.

2. Construction of the ΔmutS Δung strain

We used the EcNR1 strain as the mutS+ ung+ background and the EcNR2 strain as the *ΔmutS* background to test the targeted deamination frequency. To obtain the ΔmutS Δung background, we disrupted the ung gene in the EcNR2 strain by inserting a Zeocin resistance cassette in the middle of the gene. In brief, a Zeocin resistance cassette flanked by ung homology regions was PCR amplified from pEM7-Zeo vector (Invitrogen) using the 5'ung-zeo and 3'zeo-ung primers. This PCR fragment was transformed into the EcNR2 reporter strain. Successful disruption of ung was selected based on Zeocin resistance.

3. Protocol of MAGE and dsDNA-mediated homologous recombination

Single colonies were inoculated into LB-min media and cultured under 34 °C to an absorbance (600 nm) of 0.4~0.6. The bacterial culture was then shifted to 42 °C for 15 min to induce expression of the λ-Red recombination proteins (Exo, Beta and Gam), and then immediately chilled on ice (up to 2 hours). 1 ml of bacterial culture was centrifuged at 16,000x g for 30 s and washed twice with 1 ml dH20 at 4 °C. Cell pellets were re-suspend with 50µl DNA-containing water (100ng dsDNA fragment or 200pmole ssDNA) and transferred to a pre-chilled 1 mm gap electroporation cuvette (Bio-Rad), and electroporated with a Bio-Rad GenePulser electroporation system under the following parameters: 1.8 kV, 200 Ω and 25 µF. 1 ml S.O.C (New England Biolabs) was immediately added to the electroporated cells. The cells were recovered in S.O.C at 34 °C for 2–2.5h before plating on LB-min agar plates to resolve single colonies. The plates were incubated for at least 13h at 34 °C. Colony PCR followed by Sanger sequencing was performed to screen for colonies with the right genotypes.

E.coli cell culture and targeted deaminase activity assay

The reporter strains were electro-transformed with the plasmids coding for targeted deaminases. Single colonies were inoculated and cultured under 34°C in LB-min- media (5g NaCl, 5g yeast extract,10g tryptone in 1L ddH2O) supplemented with 100µg/mL Carbenicillin, 25µg/mL Chloramphenicol, 100µg/mL Spectinomycin, 100ug/mL Kanamycin. Targeted deaminase activities of the targeted deaminases were tested by inducing the expression the fusion protein with IPTG of final concentration 100µM when the O.D of the cell culture reached

0.4~0.6. To maintain the continuous cell proliferation, cell culture was diluted 100-fold into fresh media every 10 hours.

Flow cytometry

Targeted deaminase activity as measured by GFP+ cell fraction in the total population was assayed by flow cytometry using a LSRFortessa cell analyzer (BD Biosciences). Bacteria culture was diluted 1:100 with PBS and vortexed for 30 seconds before flow cytometry. At least 100,000 events were analyzed for each sample. Targeted deamination efficiency was calculated as the percentage of GFP positive cells in the whole population.

gfp gene Sanger sequencing

To genotype the GFP and GAPDH genes in E.coli, we inoculated single colonies in LB media and cultured them for 16 h at 34ºC. PCR reactions with Phusion enzyme (NEB) were conducted with 1µl 100X diluted bacterial culture and Sanger sequencing were performed. *Specifically, to genotype the GFP and GAPDH genes, we inoculated single col*onies in LB media and cultured them for 16 hours at 34 ºC. gfp and gapdh loci were amplified from 1ul of 100 times diluted bacterial culture using 10ul 2X Phusion High-Fidelity PCR Master Mix (NEB), 7ul water, and 1ul of 10uM primer(each) with thermocycling program of 98°C for 2 min; (98°C for 30s, 60°C for 30s, 72°C for 2min) x 30 cycles, 72°C for 10 min. Sequence of primers can be found in the Sequence 4.8.

Genomic library preparation

Corresponding reporter strains were transformed with ZFP-8aa-AID and TALE-C3-AID respectively. Single colonies were inoculated and split into the induction and non-induction groups. The expression of the deaminases was induced for 10 hours and the cell culture was plated on IPTG containing agar plate to isolate single colonies. After approximately 24h, we inoculated single colonies into LB-min media and cultured them overnight at 34ºC. In order to extract chromosomal DNA and minimize the amount of plasmid DNA, Miniprep was first performed (Qiagen) according to manufacturer's protocol and the sodium acetate/SDS precipitate formed was resuspended in the Lysis Buffer Type 2 (Illustra bacteria genomicPrep Mini Spin Kit, GE Healthcare) and the genomic DNA was recovered following manufacturer's instructions (Illustra bacteria genomicPrep Mini Spin Kit, GE Healthcare). Genomic DNA libraries were constructed from 1.5–2 µg of genomic DNA. DNA was sheared in TE buffer (10 mM Tris (pH 8.0) 0.1 mM EDTA) using microTube (Covaris) with recommended protocol. Median DNA fragment sizes, estimated by gel-electrophoresis, were 150–250 bp. Sheared fragments were processed with the DNA Sample Prep Master Mix Set 1 (NEB). Adaptors consisted of the Illumina genomic DNA adaptor oligonucleotide sequences with the addition of 2-bp barcodes. Eight barcoded genomic libraries were pooled with equal molar amount.

Specifically, the reporter strain for testing ZF-AID-targeted deamination activity was transformed with ZF-8aa-AID and one single colony was inoculated. Cells from the colony were cultured in 2mL LB-min media supplemented with 1mM IPTG to induce the expression ZF-8aa-AID. In parallel, cells from the same colony were cultured in 2mL LB-min media without induction. After 10 hours, the bacterial culture was plated on the IPTG-containing agar plate and 3 GFP+ colonies (after IPTG induction) and 1 GFP- colony (negative control without IPTG induction) were inoculated and cultured in 2mL LB-min media overnight at 34ºC. The same

work flow was undertaken for experiments with TALE-C1-AID and its corresponding reporter strain.

In order to extract chromosomal DNA and minimize the amount of plasmid DNA, Miniprep was first performed (Qiagen) on the 2mL bacteria culture according to manufacturer's protocol. The sodium acetate/SDS precipitate formed was resuspended in the Lysis Buffer Type 2 (Illustra bacteria genomicPrep Mini Spin Kit, GE Healthcare) and the genomic DNA was recovered following manufacturer's instructions. Genomic DNA libraries were constructed from 1.5–2 µg of genomic DNA. DNA was sheared in TE buffer (10 mM Tris (pH 8.0), 0.1 mM EDTA) using microTube (Covaris) with duty cycle as 10%, intensity as 5, cycle per burst as 200 and time as 780sec per sample. Median DNA fragment sizes as estimated by gel-electrophoresis, were 150–250 bp. Sheared fragments were processed with the DNA Sample Prep Master Mix Set 1 (NEB). Adaptors consisted of the Illumina genomic DNA adaptor oligonucleotide sequences with the addition of 2-bp barcodes. Eight barcoded genomic libraries were pooled with an equal molar ratio. The sequences of the adaptors and primers can be found in the Supplementary Sequence.

The sequences of the adaptors and primers can be found in the Sequence 4.9.

Genomic DNA sequencing analysis

The reference genomic sequence for the reporter strain was generated by manually modifying the FASTA sequence of E. coli K-12 strain MG1655 to reflect the removal of *mutS* and *ung*, the insertion of the lambda prophage genome into the *bioAB* operon, and the insertion of the GFP reporter into the *galK* cassette. Genomic libraries were single-end sequenced using an Illumina Genome Analyzer, generating 100bp reads. The reads were first assigned to samples

according to their 2-bp barcodes by exact matching and reads with fewer than 60 bases of high-quality sequence were discarded. Sorted reads were then aligned to the reference genomes using the Breseq package[32]. Match lengths of at least 40 bases were required for alignment. In addition to Breseq's single nucleotide variation (SNV) calling functionality, the SAMtools package[32] was used on the resulting BAM file to corroborate short indels and single nucleotide variants. To validate the result of Breseq, MAQ was used as a second method to align the raw reads to the reference genomes and to call SNVs. FastQ files containing the sequencing reads were split based on the barcode, and trimmed using the FASTX-toolkit library. The resulting fastQ files were mapped to the reference genomes with MAQ[33]. Single nucleotide substitutions were considered valid when supported by a minimum read depth of 10 or a Phred-like consensus quality higher than 80. Finally, these three sets were merged to generate the final SNV set. SNVs called by both MAQ and SAMtools, or SNVs called by one and also called by Breseq, were kept. Indels were called by SAMtools alone. Breseq was used to identify new junctions using candidates generated by split read alignment. LiftOver[34] was used to map the SNVs back to the original MG1655 genome (NCBI accession: NC_000913) for annotation. SNV effect prediction was done using the snpEff package[35] and BioPerl. The analysis flow map (Figure 4_12) provides information about the number of raw sequence reads, aligned reads, genome coverage and validated SNVs, and the list of SNVs can be found in the supplementary information.

Statistical analysis of the whole-genome sequence data

Wilcoxon test was used to analyze whether the mutation rate was higher in the strains with TALE-AID or ZF-AID induction. Intended mutations in TALE-AID, and ZF-AID strains were discarded for this analysis. X (SNVs not induced)=31, 21; Y(SNVs induced)=40, 31, 27, 33,

20, 32. H_0: There is no difference in the mutation rate; H_1: induced strains have a higher mutation rate. P_{value}=0.25. The null hypothesis cannot be rejected. Therefore, there was no significant difference in the number of mutations between induced and uninduced strains. Due to the limited sample size, sensitivity simulations were performed to ensure an appropriate type II error. 1) Random samples from the observations of size m+n were taken, and divided in two groups A (n members) and B (m members). 2) An arbitrary value δ was added to B. 3) Wilcoxon one-sided p-value was calculated for comparison of groups A and B. A p-value under 0.05 was considered a success and recorded; otherwise a failure was recorded. 4) Steps 1 to 3 were repeated 10,000 times. The estimated power of the test was approximated by the proportion of successes

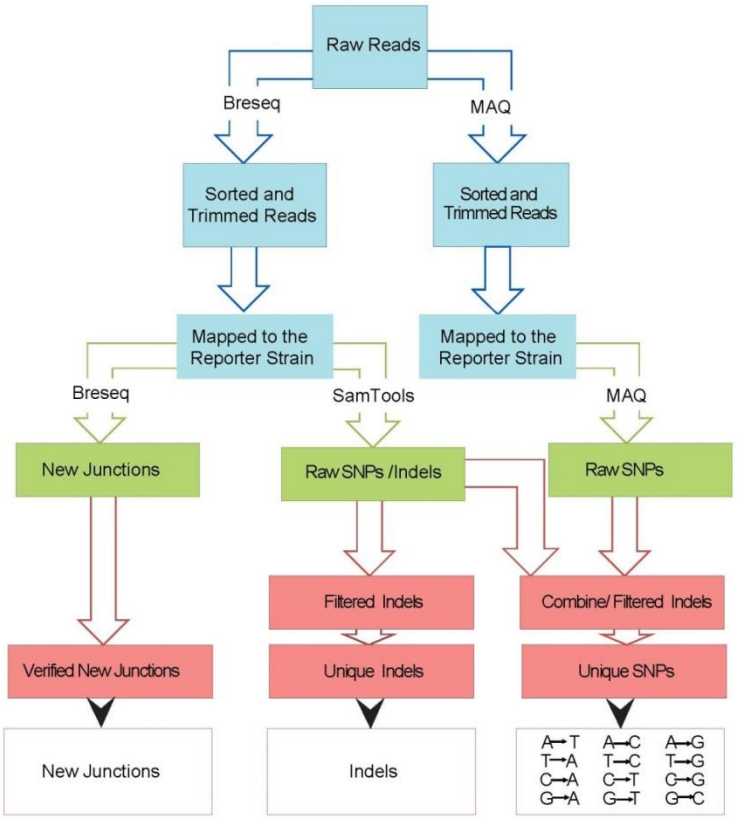


Figure 4_12 Flow map of the whole-genome sequence data analysis.
Breseq and MAQ were used independently to assign the raw reads to different strains and align the reads to the reference genomes. After alignment, we used Samtools and MAQ to identify single nucleotide substitutions (SNSs), Breseq to identify new genomic junctions and Samtools to call indels.

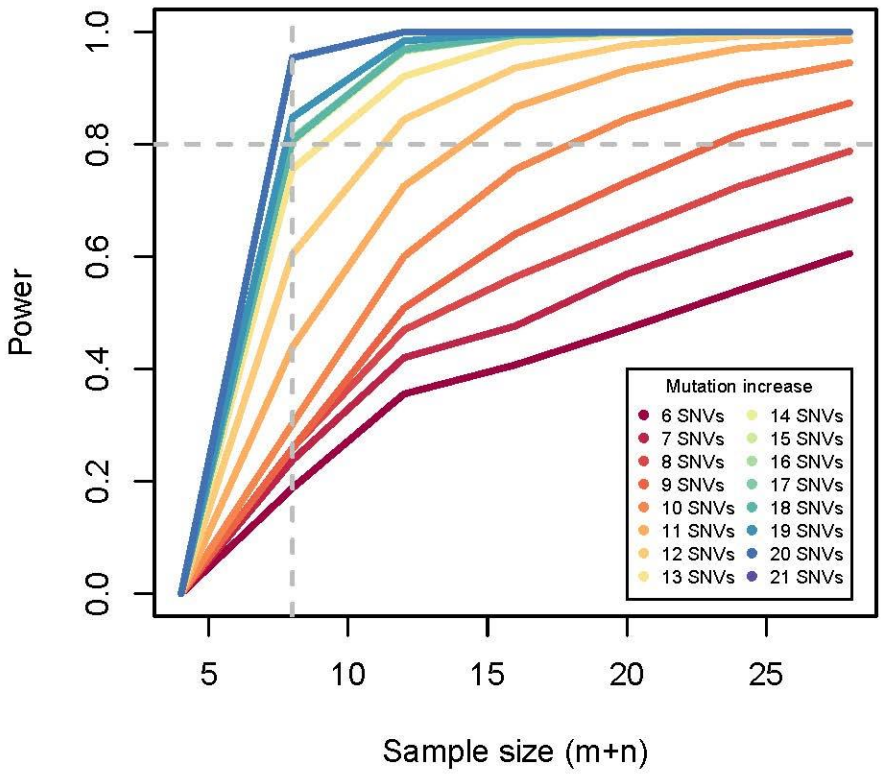


Figure 4_13 Sensitivity simulations for the Wilcoxon test of numbers of genome SNV comparison

among the 10,000 repetitions. 5) Steps 1 to 4 were repeated for a range of values of m+n and a range of δ values. The results are presented in the Figure 4_13. With the current sample size, we could detect an increase of 13 SNVs, or higher (Statistical power=0.8).

Poisson based modeling of number of genome edited sites: There are four sites in the genome with equivalent features as the targeted site. All of them contain an exact ZF binding sequence 11bp away from an upstream WRC motif. Deamination was only detected in the targeted site with a maximum frequency of 7%. Assuming that alterations of these sites are Poisson distributed with μ = .07, the probability of detecting a second mutation in any strain is 0.03, and the probability *P* of *not detecting* an additional mutation in any of the 3 ZF-AID strains is 0.90.

$$\mathrm{P}(\mathrm{k} > 1 | k \geq 1) = \frac{1 - \mathrm{P}(\mathrm{k} = 0) - \mathrm{P}(\mathrm{k} = 1)}{1 - \mathrm{P}(\mathrm{k} = 0)} = \frac{1 - \mathrm{e}^{-0.07} - 0.07\mathrm{e}^{-0.07}}{1 - \mathrm{e}^{-0.07}} = 0.03$$

$$\mathrm{P} = (1 - \mathrm{P}(\mathrm{k} > 1 | \mathrm{k} \geq 1))^3 = 0.90$$

Human cell culture

The human embryonic kidney cell line HEK293FT (Invitrogen) and the derivative reporter cell lines was maintained under 37 °C, 5% CO_2 using Dulbecco's modified Eagle's Medium supplemented with 10% FBS, 2 mM GlutaMAX (Invitrogen), 100 U/ml penicillin and 100 μg/ml streptomycin.

Targeted deaminase activity assay in human cells

The GFP-ACG reporter cell lines were generated by lentiviral transduction with low virus titration to make sure at most one copy of the reporter can be integrated into the genome. Single cells were isolated via FACS based on mCherry signal (Beckman Coulter MoFlo). Deaminase activity was tested by transfecting reporter cells with plasmids carrying ZF-AIDs. Briefly, HEK293FT cells were seeded into 12-well plates the day before transfection at densities of 4 × 105 cells/well. Approximately 24 h after initial seeding, cells were transfected using Lipofectamine 2000 (Invitrogen) and 1.6ug DNA (400 ng of ZF-AID expression plasmid and/or 20 ng of UGI expression plasmid, and/or 20ng of ShRNA-MSH6 expression plasmid (Sigma), and pUC19(Invitrogen) plasmid to 1.6ug) per well. After 48 h, cells were trypsinized from their culturing plates and resuspended in 200 μl of media for flow cytometry analysis. At least 25,000 events were analyzed for each transfection sample. The flow cytometry data were analyzed using BD FACSDiva (BD Biosciences). The reporter and constructs sequences can be found in the Sequences 4.10-4.12.

Genotyping of human cell

To genotype the GFP target locus in HEK293 cells, we picked single GFP+ monocolonies and added each to 10ul 1X prepGEM buffer and enzyme (ZyGEM). After cell lysis according to the manufacturer's instructions, the bulk product was added to a PCR reaction containing Platinum Taq polymerase (invitrogen). PCR products were cloned in pCR™4-TOPO (invitrogen) and capillary sequenced by Genewiz.

DSB generating potential assay

The GFP-In reporter[7] cell lines were generated by lentiviral transduction and successful reporter insertions were selected via puromycin selection. GFP-In reporter cells were plated in 12-well plates the day before transfection at densities of 4 × 105 cells/well transfected using Lipofectamine 2000 (Invitrogen) and 1.6ug DNA (400 ng of ZF-AID/ZF-nuclease/I-Sce1 expression plasmid and/or 20 ng of UGI expression plasmid, and/or 20ng of ShRNA-MSH6 expression plasmid, 1ug of DNA donor pUC19(Invitrogen) plasmid to 1.6ug) per well. 72 hours after the transfection, cell were trypsinized and resuspended in 200 µl of media for flow cytometry analysis. At least 25,000 events were analyzed for each transfection sample. The flow cytometry data were analyzed using BD FACSDiva (BD Biosciences). The constructs sequence can be found in the Sequences 4.13.

Cytotoxicity assay

The assays were conducted as described before[36]. Briefly, HEK293FT cells were seeded in 12-well plates ($4X10^5$ cells/well) and transfected after 24 h with 200 ng of deaminase/ nuclease expression plasmids, 10 ng of pmaxGFP (Lonzon), and pUC19 to 2 µg using calcium phosphate-mediated protocol. After 2 and 5 days, the fractions of GFP-positive cells were determined by flow cytometry (BD Biosciences). The survivability was calculated as the percentage of GFP-positive cell surviving at day 5 divided by the percentage of GFP-positive cells determined at day 2 after transfection. This ratio was normalized to the corresponding ratio after I-*Sce*I transfection, to yield the percentage survival as compared to I-*Sce*I.

Test of targeted deamination frequency on a reporter with two ZF binding sites

Although AID was observed to function as a monomer, it has also been postulated that AID forms homodimers and homotetramers based on structural modeling with homologous cytidine deaminases (*12*, *27*). Having shown that a single ZF binding site was sufficient for ZF-AID editing (Figure 4_1), we next sought to test whether we can increase the targeted deamination frequency by adding another zinc finger binding site. This would facilitate dimerziation of AID, if it functions as a dimer. To this end, we first sought to modify the reporter by adding two ZF binding sites flanking the targeting site (the broken start codon, ACG) while ensuring that the modifications would not compromise the expression of the GFP protein. Four different modified GFP reporters were investigated (Figure 4_2), however only one reporter in which an additional ZF binding sequence (5'GCCGACGTG3' in the bottom strand) lay 14bp upstream of the start codon did not compromise the translation efficiency (Figure 4_2). Therefore, we further modified this reporter by mutating its start codon to ACG with MAGE and used it to conduct further studies. Interestingly, induction of ZF-AID led to similar GFP rescue frequency (0.1%) on the dimer reporter as the one with a single ZF binding site (0.12%) (Figure 4_2), indicating that the targeting a single copy of ZF-AID at the targeting site is sufficient to exert deaminase activity in the cell. Future experiment with symmetrical zinc finger DNA binding sites is needed to substantiate this conclusion. Also, test the deamination frequency of ZF-AID with a mutated dimerization interface [1-2] might help determine the *stoichiometry* of the functional ZF-AIDs.

Sequences:

Sequence 4.1. ZF -AID constructs and the PCR primer sequences

ZF coding sequence is capitalized, linker sequence is highlighted in green, AID coding sequence is un-capitalized. NheI cutting site is labeled in Red, HindIII cutting site is labeled in Blue.

ZF -4aa-AID:

GCTAGCCCCAGAGTGAGAACCGGTTCTAAGACACCTCCCCACGAGAGGCCTTTTCAGTGTAGAATTTGTATGC
GTAATTTTTCTAGGTCCGATGTGCTGGCCAATCACACAAGGACTCACACTGGTGAAAAGCCCTTCCAATGTAG
AATTTGTATGCGCAATTTTTCTCAATCTTCTACTCTGACTAGACATCTGAGGACCCACACAGGCGAAAAGCCTT
TCCAGTGCAGAATTTGTATGAGAAATTTTTCTGAAAGACAGGGTCTGAAAAGACATCTGAAGACACATACAG
GTGAAAAAGGATCCGGTGGTGGTTCTgacagcctcttgatgaaccggaggaagtttctttaccaattcaaaaatgtccgctgggctaagggtcggcgtgaga
cctacctgtgctacgtagtgaagaggcgtgacagtgctacatccttttcactggactttggttatcttcgcaataagaacggctgccacgtggaattgctcttcctccgctacatctcggac
tgggacctagaccctggccgctgctaccgcgtcacctggttcacctcctggagcccctgctacgactgtgcccgacatgtggccgactttctgcgagggaaccccaacctcagtctg
aggatcttcaccgcgcgcctctacttctgtgaggaccgcaaggctgagcccgaggggctgcggcggctgcaccgcgccggggtgcaaatagccatcatgaccttcaaagattatttt
tactgctggaatacttttgtagaaaaccacgaaagaactttcaaagcctgggaagggctgcatgaaaattcagttcgtctctccagacagcttcggcgcatccttttgcccctgtatgagg
ttgatgacttacgagacgcatttcgtactttgggactt

ZF -4aa2-AID:

GCTAGCCCCAGAGTGAGAACCGGTTCTAAGACACCTCCCCACGAGAGGCCTTTTCAGTGTAGAATTTGTATGC
GTAATTTTTCTAGGTCCGATGTGCTGGCCAATCACACAAGGACTCACACTGGTGAAAAGCCCTTCCAATGTAG
AATTTGTATGCGCAATTTTTCTCAATCTTCTACTCTGACTAGACATCTGAGGACCCACACAGGCGAAAAGCCTT
TCCAGTGCAGAATTTGTATGAGAAATTTTTCTGAAAGACAGGGTCTGAAAAGACATCTGAAGACACATACAG
GTGAAAAAGGATCCCTGCGTGGTTCTgacagcctcttgatgaaccggaggaagtttctttaccaattcaaaaatgtccgctgggctaagggtcggcgtgaga
cctacctgtgctacgtagtgaagaggcgtgacagtgctacatccttttcactggactttggttatcttcgcaataagaacggctgccacgtggaattgctcttcctccgctacatctcggac
tgggacctagaccctggccgctgctaccgcgtcacctggttcacctcctggagcccctgctacgactgtgcccgacatgtggccgactttctgcgagggaaccccaacctcagtctg
aggatcttcaccgcgcgcctctacttctgtgaggaccgcaaggctgagcccgaggggctgcggcggctgcaccgcgccggggtgcaaatagccatcatgaccttcaaagattatttt
tactgctggaatacttttgtagaaaaccacgaaagaactttcaaagcctgggaagggctgcatgaaaattcagttcgtctctccagacagcttcggcgcatccttttgcccctgtatgagg
ttgatgacttacgagacgcatttcgtactttgggactt

ZF-8AA-AID

GCTAGCCCCAGAGTGAGAACCGGTTCTAAGACACCTCCCCACGAGAGGCCTTTTCAGTGTAGAATTTGTATGC
GTAATTTTTCTAGGTCCGATGTGCTGGCCAATCACACAAGGACTCACACTGGTGAAAAGCCCTTCCAATGTAG
AATTTGTATGCGCAATTTTTCTCAATCTTCTACTCTGACTAGACATCTGAGGACCCACACAGGCGAAAAGCCTT
TCCAGTGCAGAATTTGTATGAGAAATTTTTCTGAAAGACAGGGTCTGAAAAGACATCTGAAGACACATACAG
GTGAAAAAGGATCCTCTGGTGGTGGTCTGGGGTCGACTgacagcctcttgatgaaccggaggaagtttctttaccaattcaaaaatgtccgctgg
gctaagggtcggcgtgagacctacctgtgctacgtagtgaagaggcgtgacagtgctacatccttttcactggactttggttatcttcgcaataagaacggctgccacgtggaattgctc
ttcctccgctacatctcggactgggacctagaccctggccgctgctaccgcgtcacctggttcacctcctggagcccctgctacgactgtgcccgacatgtggccgactttctgcgag
ggaaccccaacctcagtctgaggatcttcaccgcgcgcctctacttctgtgaggaccgcaaggctgagcccgaggggctgcggcggctgcaccgcgccggggtgcaaatagcca
tcatgaccttcaaagattatttttactgctggaatacttttgtagaaaaccacgaaagaactttcaaagcctgggaagggctgcatgaaaattcagttcgtctctccagacagcttcggcgc
atccttttgcccctgtatgaggttgatgacttacgagacgcatttcgtactttgggactt

ZF-11AA-AID:

GGTTCTGGTGGTGGTGGTTCTGGTGGTGGTGGTGCTAGCCCCAGAGTGAGAACCGGTTCTAAGACACCTCCCC
ACGAGAGGCCTTTTCAGTGTAGAATTTGTATGCGTAATTTTTCTAGGTCCGATGTGCTGGCCAATCACACAAG
GACTCACACTGGTGAAAAGCCCTTCCAATGTAGAATTTGTATGCGCAATTTTTCTCAATCTTCTACTCTGACTA
GACATCTGAGGACCCACACAGGCGAAAAGCCTTTCCAGTGCAGAATTTGTATGAGAAATTTTTCTGAAAGACA
GGGTCTGAAAAGACATCTGAAGACACATACAGGTGAAAAAGGATCCGGTTCTGGTGGTGGTGGTTCTGGTGG
TGGTGGTgacagcctcttgatgaaccggaggaagtttctttaccaattcaaaaatgtccgctgggctaagggtcggcgtgagacctacctgtgctacgtagtgaagaggcgtga
cagtgctacatccttttcactggactttggttatcttcgcaataagaacggctgccacgtggaattgctcttcctccgctacatctcggactgggacctagaccctggccgctgctaccgc
gtcacctggttcacctcctggagcccctgctacgactgtgcccgacatgtggccgactttctgcgagggaaccccaacctcagtctgaggatcttcaccgcgcgcctctacttctgtga
ggaccgcaaggctgagcccgaggggctgcggcggctgcaccgcgccggggtgcaaatagccatcatgaccttcaaagattatttttactgctggaatacttttgtagaaaaccacga
agaactttcaaagcctgggaagggctgcatgaaaattcagttcgtctctccagacagcttcggcgcatccttttgcccctgtatgaggttgatgacttacgagacgcatttcgtacttg
ggactt

Primers for ZF-AID constructs

ZFP-F	ATCGGCTAGCCCCAGAGTGAGAACCGGT
ZFP-R-4AA	ccggttcatcaagaggctgtcAGAACCACCACCGGATCCTTTTTCACCTGTATG
ZFP-R-4AA2	ccggttcatcaagaggctgtcAGAACCACGCAGGGATCCTTTTTCACCTGTATG
ZFP-R-8AA	ctgtcAGTCGACCCCAGACCACCACCAGAGGATCCTTTTTCACCTGTATG
ZFP-R-11AA	gtcACCACCACCACCAGAACCACCACCACCAGAACCGGATCCTTTTTCACCTGTATG
ZFP-R-Hind3	atcgaagcttGGATCCTTTTTCACCTGTATG
AID-F-Nhe1	ATCGGCTAGCgacagcctcttgatgaaccg
AID-F-4AA	TACAGGTGAAAAAGGATCCGGTGGTGGTTCTgacagcctcttgatgaaccg
AID-F-4AA2	TACAGGTGAAAAAGGATCCCTGCGTGGTTCTgacagcctcttgatgaaccg
AID-F-8AA	AGGATCCTCTGGTGGTGGTCTGGGGTCGACTgacagcctcttgatgaaccg
AID-F-11AA	GATCCGGTTCTGGTGGTGGTGGTTCTGGTGGTGGTGGTgacagcctcttgatgaaccg
AID-R	atcgaagcttaaagtcccaaagtacgaaatgcg

Sequence 4.2 TALE-AID constructs and the PCR primer sequences

TALE-AID full sequence: TALE N-terminus is in Blue , TALE central repeating domain is in Red, TALE-C terminus is in Green. Linker sequence is in Brown. AID coding sequence is un-capitalized in Black.

ATGTCGCGGACCCGGCTCCCTTCCCCACCCGCACCCAGCCCAGCGTTTTCGGCCGACTCGTTCTCAGACCTGCT
TAGGCAGTTCGACCCCTCACTGTTTAACACATCGTTGTTCGACTCCCTTCCTCCGTTTGGGGCGCACCATACGG
AGGCGGCCACCGGGGAGTGGGATGAGGTGCAGTCGGGATTGAGAGCTGCGGATGCACCACCCCCAACCATGC
GGGTGGCCGTCACCGCTGCCCGACCGCCGAGGGCGAAGCCCGCACCAAGGCGGAGGGCAGCGCAACCGTCC
GACGCAAGCCCCGCAGCGCAAGTAGATTTGAGAACTTTGGGATATTCACAGCAGCAGCAGGAAAAGATCAAG
CCCAAAGTGAGGTCGACAGTCGCGCAGCATCACGAAGCGCTGGTGGGTCATGGGTTTACACATGCCCACATC
GTAGCCTTGTCGCAGCACCCTGCAGCCCTTGGCACGGTCGCCGTCAAGTACCAGGACATGATTGCGGCGTTGC
CGGAAGCCACACATGAGGCGATCGTCGGTGTGGGGAAACAGTGGAGCGGAGCCCGAGCGCTTGAGGCCCTGT
TGACGGTCGCGGGAGAGCTGAGAGGGCCTCCCCTTCAGCTGGACACGGGCCAGTTGCTGAAGATCGCGAAGC
GGGGAGGAGTCACGGCGGTCGAGGCGGTGCACGCGTGGCGCAATGCGCTCACGGGAGCACCCCTCAACCTGA
CCCCAGAGCAGGTCGTGGCAATCGCCTCCAACATTGGCGGGAAACAGGCACTCGAGACTGTCCAGCGCCTGC
TTCCCGTGCTGTGCCAAGCGCACGGACTCACCCCAGAGCAGGTCGTGGCGATCGCAAGCCACGACGGAGGAA
AGCAAGCCTTGGAAACAGTACAGAGGCTGTTGCCTGTGCTGTGCCAAGCGCACGGCCTCACCCCAGAGCAGG
TCGTGGCAATCGCGAGCAATAACGGCGGAAAACAGGCTTTGGAAACGGTGCAGAGGCTCCTTCCAGTGCTGT
GCCAAGCGCACGGATTAACCCCAGAGCAGGTCGTGGCAATCGCCTCCAACATTGGCGGGAAACAGGCACTCG
AGACTGTCCAGCGCCTGCTTCCCGTGCTGTGCCAAGCGCACGGCTTAACCCCAGAGCAGGTCGTGGCGATCGC
AAGCCACGACGGAGGAAAGCAAGCCTTGGAAACAGTACAGAGGCTGTTGCCTGTGCTGTGCCAAGCGCACGG
ACTTACCCCAGAGCAGGTCGTGGCCATTGCCTCGAATGGAGGGGGCAAACAGGCGTTGGAAACCGTACAACG
ATTGCTGCCGGTGCTGTGCCAAGCGCACGGCCTTACCCCAGAGCAGGTCGTGGCGATCGCAAGCCACGACGG
AGGAAAGCAAGCCTTGGAAACAGTACAGAGGCTGTTGCCTGTGCTGTGCCAAGCGCACGGACTAACCCCAGA
GCAGGTCGTGGCAATCGCCTCCAACATTGGCGGGAAACAGGCACTCGAGACTGTCCAGCGCCTGCTTCCCGTG
CTGTGCCAAGCGCACGGGCTCACCCCAGAGCAGGTCGTGGCGATCGCAAGCCACGACGGAGGAAAGCAAGC
CTTGGAAACAGTACAGAGGCTGTTGCCTGTGCTGTGCCAAGCGCACGGGCTAACCCCAGAGCAGGTCGTGGC
CATTGCCTCGAATGGAGGGGGCAAACAGGCGTTGGAAACCGTACAACGATTGCTGCCGGTGCTGTGCCAAGC
GCACGGCCTAACCCCAGAGCAGGTCGTGGCAATCGCCTCCAACATTGGCGGGAAACAGGCACTCGAGACTGT
CCAGCGCCTGCTTCCCGTGCTGTGCCAAGCGCACGGGTTAACCCCAGAGCAGGTCGTGGCCATTGCCTCGAAT
GGAGGGGGCAAACAGGCGTTGGAAACCGTACAACGATTGCTGCCGGTGCTGTGCCAAGCGCACGGACTCACG
CCTGAGCAGGTAGTGGCTATTGCATCCAATATCGGGGGCAGACCCGCACTGGAGTCAATCGTGGCCCAGCTTT
CGAGGCCGGACCCCGCGCTGGCCGCACTCACTAATGATCATCTTGTAGCGCTGGCCTGCCTCGGCGGACGACC
CGCCTTGGATGCGGTGAAGAAGGGGCTCCCGCACGCGCCTGCATTGATTAAGCGGACCAACAGAAGGATTCC
CGAGAGGACATCACATCGAGTGGCAGATCACGCGCAAGTGGTCCGCGTGCTCGGATTCTTCCAGTGTCACTCC
CACCCCGCACAAGCGTTCGATGACGCCATGACTCAATTTGGTATGTCGAGACACGGACTGCTGCAGCTCTTTC
GTAGAGTCGGTGTCACAGAACTCGAGGCCCGCTCGGGCACACTGCCTCCCGCCTCCCAGCGGTGGGACAGGA
TTCTCCAAGCGAGCGGTATGAAACGCGCGAAGCCTTCACCTACGTCAACTCAGACACCTGACCAGGCGAGCC
TTCATGCGTTCGCAGACTCGCTGGAGAGGGATTTGGACGCGCCCTCGCCCATGCATGAAGGGGACCAAACTC
GCGCGTCAGCTAGCCCCAAGAAGAAGAGAAAGGTGGAGGCCAGCgacagcctcttgatgaaccggaggaagtttctttaccaattcaaaa
atgtccgctgggctaagggtcggcgtgagacctacctgtgctacgtagtgaagaggcgtgacagtgctacatccttttcactggactttggttatcttcgcaataagaacggctgccac
gtggaattgctcttcctccgctacatctcggactgggacctagaccctggccgctgctaccgcgtcacctggttcacctcctggagcccctgctacgactgtgcccgacatgtggccg
actttctgcgagggaaccccaacctcagtctgaggatcttcaccgcgcgcctctacttctgtgaggaccgcaaggctgagcccgaggggctgcggcggctgcaccgcgccggggt

gcaaatagccatcatgaccttcaaagattatttttactgctggaatacttttgtagaaaaccacgaaagaactttcaaagcctgggaagggctgcatgaaaattcagttcgtctctccagac
agcttcggcgcatccttttgcccctgtatgaggttgatgacttacgagacgcatttcgtactttgggactttga

Primer sequences

TAL-F-ssp1	TGGCAAATATTCTGAAATGAGCTGTTGACAATTAATCATCCGGTCCGTATAATCT GTGGAATTGTGAGCGGATAACAATTTCACACAAAGAGGAGAAAGGTACCATGT CGCGGACCCGGCTCCC
TAL-C1-Nhe1	TGGGGCTAGCTGACGCGCGAGTTTGGTCCC
TAL-C2-Nhe1	TCTTGGGGCTAGCGCGGGAGGCAGTGTGCCCGA
TAL-C3-Nhe1	TCTTGGGGCTAGCTGCCACTCGATGTGATGTCCTCTCGGGAATCCT
TAL-C4-Nhe1	TCTTGGGGCTAGCGCGGCCAGCGCGGGGTCCG
TAL-C5-Nhe1	TCTTGGGGCTAGCCTCCAGTGCGGGTCTGCC
AID-R	atcgaagcttaaagtcccaaagtacgaaatgcg

Sequence 4.3 APOBECs and PCR primer sequences

APOBEC1 sequence:

ACTTCTGAAAAAGGTCCATCTACTGGTGATCCTACTCTGCGTCGTCGTATTGAACCGTGGGAATTTGACGTGTT
CTACGACCCACGCGAACTGCGTAAAGAGGCTTGCCTGCTGTACGAAATCAAATGGGGTATGTCTCGCAAAATT
TGGCGCTCCAGCGGTAAAAACACCACTAACCACGTTGAAGTCAACTTCATCAAAAAGTTCACCTCTGAACGCG
ACTTCCACCCGTCCATGTCTTGTTCTATCACCTGGTTCCTGTCTTGGAGCCCGTGCTGGGAGTGCTCCCAAGCC
ATCCGCGAATTCCTGTCTCGTCACCCGGGTGTAACGCTGGTGATCTATGTCGCCCGTCTGTTCTGGCATATGGA
TCAGCAAAACCGTCAGGGTCTGCGTGATCTGGTGAACAGCGGCGTCACGATCCAGATCATGCGTGCATCCGA
ATATTACCATTGCTGGCGTAACTTCGTAAACTACCCTCCGGGTGATGAAGCGCACTGGCCGCAATACCCGCCG
CTGTGGATGATGCTGTACGCTCTGGAGCTGCATTGCATCATCCTGTCTCTGCCACCGTGCCTGAAAATTTCCCG
CCGTTGGCAGAACCATCTGACCTTCTTCCGTCTGCATCTGCAGAACTGTCACTACCAGACTATCCCGCCTCACA
TCCTGCTGGCTACTGGCCTGATCCATCCGTCTGTTGCGTGGCGC

APOBEC3F sequence

AAACCGCATTTTCGTAACACCGTTGAGCGTATGTATCGTGACACTTTCTCTTACAACTTCTACAACCGTCCGAT
CCTGTCTCGCCGCAACACCGTGTGGCTGTGTTATGAAGTTAAAACCAAAGGCCCGTCTCGTCCGCGTCTGGAC
GCGAAGATCTTCCGTGGCCAGGTACCGCGTTCCTTTATTCGTGCGCCGTTTCAGGTGCTGTCTAGCCCGTTCGG
CCAGTGTGCACCGCCGCACGGTACGGCGCAGGTTCAATGGCCTCCGCAGCTGACTGCCGGTCGCGAGCAGGG
TCGTCCG

APOBEC3G 2K3A sequence

GAAATTCTGCGTCACTCTATGGACCCGCCAACTTTTACTTTCAACTTCAACAATGAACCGTGGGTCCGTGGCCG
TCACGAGACTTACCTGTGCTACGAGGTGGAGCGTATGCACAATGATACCTGGGTGAAACTGAACCAGCGTCG
CGGTTTCCTGGCTAACCAGGCTCCGCACAAACACGGCTTCCTGGAGGGCCGTCACGCTGAACTGTGCTTCCTG
GATGTTATTCCTTTCTGGAAACTGGACCTGGACCAAGATTATCGTGTAACTTGCTTCACTAGCTGGAGCCCATG
CTTCAGCTGCGCACAGGAAATGGCCAAGTTCATTTCTAAAAACAAACATGTTTCTCTGTGTATCAAGACTGCT
CGCATCTATGATGACCAGGGCCGTGCTCAGGAAGGCCTGCGTACTCTGGCGGAAGCAGGTGCTAAAATTAGC
ATCATGACTTACAGCGAATTCAAACACTGCTGGGACACCTTCGTGGACCACCAGGGTGCGCCTTTCCAGCCTT
GGGATGGTCTGGATGAACACTCTCAGGACCTGTCTGGTCGTCTGCGTGCGATCCTGCAGAACCAGGAAAAT

Primers for ZF-APOBECs constructs

Homology to the vectors is in Black; linker sequence is highlighted in Green and the homology to the APOBECs sequences is in Red.

APOBEC1-F-4AA1	GAAAAGACATCTGAAGACACATACAGGTGAAAAAGGATCCGGTGGTGGTTCTA CTTCTGAAAAAGGTCCATCTAC
APOBEC1-reverse	CCATGGGATCCCCCGGGCTGCAGGAATTCGATATCAAGCTTTCAGCGCCACGCA ACAGAC
APOBEC3F-F-4AA1	GAAAAGACATCTGAAGACACATACAGGTGAAAAAGGATCCGGTGGTGGTTCT AAACCGCATTTTCGTAACACCGTTGAGCG

APOBEC3F-reverse	CCATGGGATCCCCCGGGCTGCAGGAATTCGATATCAAGCTTCACGGACGACCCT GCTCGC
APOBEC3G-F-4AA1	GAAAAGACATCTGAAGACACATACAGGTGAAAAAGGATCCGGTGGTGGTTCT GAAATTCTGCGTCACTCTATGGAC
APOBEC3G-reverse	CCATGGGATCCCCCGGGCTGCAGGAATTCGATATCAAGCTTTCAATTTTCCTGGT TCTGC

Primers for TALE-APOBECs construct

NheI cutting site is in Red and HindIII cutting site is in the Blue.

APOBEC1-F	ATCGGCTAGCCCCAAGAAGAAGAGAAAGGTGGAGGCCAGCACTTCTGAAAAAG GTCCATCTACTGGTG
APOBEC1-R	ATCGAAGCTTTCAGCGCCACGCAACAGACGGATGG
APOBEC3F-F	ATCGGCTAGCCCCAAGAAGAAGAGAAAGGTGGAGGCCAGCAAACCGCATTTTC GTAACACCGTTGAG
APOBEC3F-R	ATCGAAGCTTTCACGGACGACCCTGCTCGCGACCG
APOBEC-3G-F	ATCGGCTAGCCCCAAGAAGAAGAGAAAGGTGGAGGCCAGCGAAATTCTGCGTC ACTCTATG
APOBEC-3G-R	ATCGAAGCTTTCAATTTTCCTGGTTCTGCAGGATCG

Sequence 4.4 pTrc-Kan backbone sequence

CTCGAGGTGGTGAATGTGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCG
TTTCCCGCGTGGTGAACCAGGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGGAAGCGGCGATGGCGG
AGCTGAATTACATTCCCAACCGCGTGGCACAACAACTGGCGGGCAAACAGTCGTTGCTGATTGGCGTTGCCAC
CTCCAGTCTGGCCCTGCACGCGCCGTCGCAAATTGTCGCGGCGATTAAATCTCGCGCCGATCAACTGGGTGCC
AGCGTGGTGGTGTCGATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCGGTGCACAATCTTCTCGCG
CAACGCGTCAGTGGGCTGATCATTAACTATCCGCTGGATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGCA
CTAATGTTCCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCATCAACAGTATTATTTTCTCCCATGAAGAC
GGTACGCGACTGGGCGTGGAGCATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTAGCGGGCCCATTAA
GTTCTGTCTCGGCGCGTCTGCGTCTGGCTGGCTGGCATAAATATCTCACTCGCAATCAAATTCAGCCGATAGC
GGAACGGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAACCATGCAAATGCTGAATGAGGGCATCGT
TCCCACTGCGATGCTGGTTGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCGAGTCCGGGCTG
CGCGTTGGTGCGGATATCTCGGTAGTGGGATACGACGATACCGAAGACAGCTCATGTTATATCCCGCCGTTAA
CCACCATCAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACCGCTTGCTGCAACTCTCTCAGGGCCA
GGCGGTGAAGGGCAATCAGCTGTTGCCCGTCTCACTGGTGAAAAGAAAAACCACCCTGGCACCCAATACGCA
AACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGG
CAGTGAGCGCAACGCAATTAATGTGAGTTAGCGCGAATTGATCTGGTTTGACAGCTTATCATCGACTGCACGG
TGCACCAATGCTTCTGGCGTCAGGCAGCCATCGGAAGCTGTGGTATGGCTGTGCAGGTCGTAAATCACTGCAT
AATTCGTGTCGCTCAAGGCGCACTCCCGTTCTGGATAATGTTTTTTGCGCCGACATCATAACGGTTCTGGCAAA
TATTCTGAAATGAGCTGTTGACAATTAATCATCCGGTCCGTATAATCTGTGGAATTGTGAGCGGATAACAATT
TCACACAGGAAACAGACCATGTCGTACTACCATCACCATCACCATCACGATTACGATATCCCAACGACCGAA
AACCTGTATTTTCAGGGCGCCGCTAGCCCCAGCGACTAAGCTTGATATCGAATTCCTGCAGCCCGGGGGATCC
CATGGTACGCGTGCTAGAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTG
TTGTTTGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCGCCCTAGACCTAGGGCGTTCGGCTGCGGCGA
GCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATG
TGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGC
CCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATAC
CAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGC
CTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTC
GCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCT
TGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAG
GTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTATTTGGT
ATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCG
CTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTT
GATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGACTAGTGCT
TGGATTCTCACCAATAAAAAACGCCCGGCGGCAACCGAGCGTTCTGAACAAATCCAGATGGAGTTCTGAGGT
CATTACTGGATCTATCAACAGGAGTCCAAGCGAGCTCTCGAACCCCAGAGTCCCGCTCAGAAGAACTCGTCA
AGAAGGCGATAGAAGGCGATGCGCTGCGAATCGGGAGCGGCGATACCGTAAAGCACGAGGAAGCGGTCAGC
CCATTCGCCGCCAAGCTCTTCAGCAATATCACGGGTAGCCAACGCTATGTCCTGATAGCGGTCCGCCACACCC
AGCCGGCCACAGTCGATGAATCCAGAAAAGCGGCCATTTTCCACCATGATATTCGGCAAGCAGGCATCGCCA

TGGGTCACGACGAGATCCTCGCCGTCGGGCATGCGCGCCTTGAGCCTGGCGAACAGTTCGGCTGGCGCGAGC
CCCTGATGCTCTTCGTCCAGATCATCCTGATCGACAAGACCGGCTTCCATCCGAGTACGTGCTCGCTCGATGC
GATGTTTCGCTTGGTGGTCGAATGGGCAGGTAGCCGGATCAAGCGTATGCAGCCGCCGCATTGCATCAGCCAT
GATGGATACTTTCTCGGCAGGAGCAAGGTGAGATGACAGGAGATCCTGCCCCGGCACTTCGCCCAATAGCAG
CCAGTCCCTTCCCGCTTCAGTGACAACGTCGAGCACAGCTGCGCAAGGAACGCCCGTCGTGGCCAGCCACGAT
AGCCGCGCTGCCTCGTCCTGCAGTTCATTCAGGGCACCGGACAGGTCGGTCTTGACAAAAAGAACCGGGCGC
CCCTGCGCTGACAGCCGGAACACGGCGGCATCAGAGCAGCCGATTGTCTGTTGTGCCCAGTCATAGCCGAAT
AGCCTCTCCACCCAAGCGGCCGGAGAACCTGCGTGCAATCCATCTTGTTCAATCATGCGAAACGATCCTCATC
CTGTCTCTTGATCAGATCTTGAT

Sequence 4.5 pL-tetO promoter and PCR primers

CTCGAGTCCCTATCAGTGATAGAGATTGACATCCCTATCAGTGATAGAGATACTGAGCACATCAGCAGGACGC
ACTGACCGAA TTCATTAAAGAGGAGAAAGGTACC

pL-tetO-5	ATCGCTCGAGTCCCTATCAGTGATAGAGATTGACATCC
pL-tetO-3	ACTCTGGGGCTAgcCATGGTACCTTTCTCCTCTTTAATG

Sequence 4.6 GFP reporter cassette, PCR primer sequences and recombineering oligos

GFP reporter cassette

Start codon is in Blue and the ZFP binding site is highlighted in Yellow, GFP coding sequence is in Green.

CGCGAAATTAATACGACTCACTATAGGGAGACCACAACGGTTTCCCTCTAGAAATAATTTTGTTTAACTTTAA
GAAGGAGATATACATATGCGGGGTTCTGCCGCAGTGGGTATGGCTAGCATGACTGGTGGACAGCAAATGGGT
CGGGATCTGTACGACGATGACGATAAGGATCGATGGGGATCCGAATTCGCCACCATGGTGAGCAAGGGCGAG
GAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTG
TCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTG
CCCGTGCCCTGGCCCACCCTCGTGACCACCTTGACCTACGGCGTGCAGTGCTTCGCCCGCTACCCCGACCACA
TGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGA
CGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAA
GGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAAGGT
CTATATCACCGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGACCCGCCACAACATCGAGGACGG
CAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAA
CCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGA
GTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTAACTCGAGAAGCTTGATCCGGCT
GCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGG
GCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATCTGGCGT

Primers for reporter integration

5'-galk-gfp	atcaaaccgtgatcagttgtgcaccacgcgatgaccgtaaCGCGAAATTAATACGACTCAC
3'-gfp-galk	gtcgagctgattttcataatcggctgccatcacgcgaactACGCCAGATCCGGATATAGTTC

Oligo designed for reporter modification

The start codon position (ACG/ATG) is in Red, and the ZF/TALE binding site is highlighted in yellow and blue respectively. * is the phosphothioester bond.

ZFP-ACG	C*T*CTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACAAACGCGGGG TTCTGCCGCAGTGGGTATGGCTAGCATGACTGGTGGA*C*A
TAL-ACG-3bp-spacer	TAATTTTGTTTAACTTTAAGAAGGAGATATACATACGAcGGGAAGAATCGTGA GTATGGCTAGCATGACTGGTGGACAGCAAATGGGTCG
TAL-ACG-6bp-spacer	TTTGTTTAACTTTAAGAAGGAGATATACATACGATTAGGGGAAGAATCGTGAG TATGGCTAGCATGACTGGTGGACAGCAAATGGGT
TAL-ACG-9bp-spacer	TTTGTTTAACTTTAAGAAGGAGATATACATACGATTAGTCTGGGAAGAATCGT GAGTATGGCTAGCATGACTGGTGGACAGCAAATGGGT

TAL-ACG-12bp-spacer	ACTTTAAGAAGGAGATATACATACGATTAGTCTGTTGGGAAGAATCGTGAGT ATGGCTAGCATGACTGGTGGACAGCAAATGGGTCGGGA
TAL-ACG-15bp-spacer	TTGTTTAACTTTAAGAAGGAGATATACATACGATTAGTCTGTTTACGGGAAGA ATCGTGAGTATGGCTAGCATGACTGGTGGACAGCAAA
APOBEC1-ACG-ZFP	C*T*C*TAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACAAACGAAAC AACAAGCCGCAGTGGGTATGGCTAGCATGACTGGTG*G*A*C
APOBEC3F-ACG-ZFP	C*T*C*TAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATACGAAACA ACAAGCCGCAGTGGGTATGGCTAGCATGACTGGTG*G*A*C
APOBEC3G-ACG-ZFP	C*T*C*TAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACACACGCAACA ACAAGCCGCAGTGGGTATGGCTAGCATGACTGGTG*G*A*C
APOBEC1-ACG-TAL	C*C*TCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACAAACGATTA GTCTGGGAAGAATCGTGAGTATGGCTAGCATGACTGG*T*G
APOBEC3F-ACG-TAL	C*C*TCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATACGATTAG TCTGGGAAGAATCGTGAGTATGGCTAGCATGACTGG*T*G
APOBEC3G-ACG-TAL	C*T*CTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACACACGCTTAGT CTGGGAAGAATCGTGAGTATGGCTAGCATGACTGG*T*G
ATG-NNCCAA-ZFP	A*A*TTTTGTTTAACTTTAAGAAGGAGATATACAAATGANNCAATTATTACTGC CGCAGTGTGGTATGGCTAGCATGACTGGTGGACAGC*A*A

Sequence 4.7 Zeocin resistance cassette and PCR primer sequences

TTTGCTGGCCTTTTGCTCACATGTGTGCTGGGCCCAGCCGGCCAGATCTGAGCTCGCGGCCGCGATATCGCTA
GCTCGAGCACGTGTTGACAATTAATCATCGGCATAGTATATCGGCATAGTATAATACGACAAGGTGAGGAACT
AAACCATGGCCAAGTTGACCAGTGCCGTTCCGGTGCTCACCGCGCGCGACGTCGCCGGAGCGGTCGAGTTCTG
GACCGACCGGCTCGGGTTCTCCCGGGACTTCGTGGAGGACGACTTCGCCGGTGTGGTCCGGGACGACGTGAC
CCTGTTCATCAGCGCGGTCCAGGACCAGGTGGTGCCGGACAACACCCTGGCCTGGGTGTGGGTGCGCGGCCT
GGACGAGCTGTACGCCGAGTGGTCGGAGGTCGTGTCCACGAACTTCCGGGACGCCTCCGGGCCGGCCATGAC
CGAGATCGGCGAGCAGCCGTGGGGGCGGGAGTTCGCCCTGCGCGACCCGGCCGGCAACTGCGTGCACTTCGT
GGCCGAGGAGCAGGACTGAGAATTCCCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTTGGCACTG
GCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCC
CTTTCGCCAGCTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGCGCAGCCTGAATGG
CGAATGGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCACTCT
CAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGCCCCGACACCCGCCAACACCCGC

5’ung-zeo	ATGGCTAACGAATTAACCTGGCATGACGTGCTGGCTGAAGCTTTTGCTGGCCTTT TGCTC
3’zeo-ung	TTACTCACTCTCTGCCGGTAATACTGGCATCCAGTCAATCGTCAGCGGGTGTTGG CGGGT

Sequence 4.8 GFP and GAPDH amplification and sequence primer sequences

Amplification-GFP-5	Cgtttgcgcgcagtcagcgatatccattttcgcgaatccg
Ampliciation-GFP-3	CGCAGTTACAGCCTACAAACTGGTTTTCTGCTTC
Sequencing-GFP-f	Atgagtctgaaagaaaaaacacaatc
Sequencing-GFP-r	TGACCGTTAAGCGCGATTTG
Amplification-GAPDH-5	Tatttacagtcttaatgagtgaaagaggcggagg
Amplification-GAPDH-3	Gccatcctggtctaagcttggaaagg
Sequencing-GAPDH-f	Aggcggaggttttttcctccgcctgtgcgcg
Sequencing-GAPDH-r	Atcaattttcatccgaacgttcc

Sequence 4.9 Next generation adaptor and PCR primer sequences

Adaptor1	PE-A1-F	TACACTCTTTCCCTACACGACGCTCTTCCGATCTac*T
	PE-A1-R	/5Phos/gtAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
Adaptor2	PE-A2-F	TACACTCTTTCCCTACACGACGCTCTTCCGATCTtg*T
	PE-A2-R	/5Phos/caAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
Adaptor3	PE-A3-F	TACACTCTTTCCCTACACGACGCTCTTCCGATCT tc*T
	PE-A3-R	/5Phos/gaAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
Adaptor4	PE-A4-F	TACACTCTTTCCCTACACGACGCTCTTCCGATCT ga*T
	PE-A4-R	/5Phos/tcAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
Adaptor5	PE-A5-F	TACACTCTTTCCCTACACGACGCTCTTCCGATCTag*T
	PE-A5-R	/5Phos/ctAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
Adaptor6	PE-A6-F	TACACTCTTTCCCTACACGACGCTCTTCCGATCTgt*T
	PE-A6-R	/5Phos/acAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
Adaptor7	PE-A7-F	TACACTCTTTCCCTACACGACGCTCTTCCGATCT ct*T
	PE-A7-R	/5Phos/agAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
Adaptor8	PE-A8-F	TACACTCTTTCCCTACACGACGCTCTTCCGATCT ca*T
	PE-A8-R	/5Phos/tgAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
PCR primers	PE-PCR-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACG AC
	PE-PCR-2	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCT GAACC

4. Sequence 4.10 Human GFP-ACG reporter sequence and genotyping primers

The pEF-1α promoter sequence is in Blue, the GFP ORF is in Green and the IRES is highlighted in Gray and the mcherry ORF is in Red. Of note, the barcode sequence is highlighted in Yellow.

TGCAAAGATGGATAAAGTTTTAAACAGAGAGGAATCTTTGCAGCTAATGGACCTTCTAGGTCTTGAAAGGAGT
GGGAATTGGCTCCGGTGCCCGTCAGTGGGCAGAGCGCACATCGCCCACAGTCCCCGAGAAGTTGGGGGGAGG
GGTCGGCAATTGAACCGGTGCCTAGAGAAGGTGGCGCGGGGTAAACTGGGAAAGTGATGTCGTGTACTGGCT
CCGCCTTTTTCCCGAGGGTGGGGGAGAACCGTATATAAGTGCAGTAGTCGCCGTGAACGTTCTTTTTCGCAAC
GGGTTTGCCGCCAGAACACAGGTAAGTGCCGTGTGTGGTTCCCGCGGGCCTGGCCTCTTTACGGGTTATGGCC
CTTGCGTGCCTTGAATTACTTCCACTGGCTGCAGTACGTGATTCTTGATCCCGAGCTTCGGGTTGGAAGTGGGT
GGGAGAGTTCGAGGCCTTGCGCTTAAGGAGCCCCTTCGCCTCGTGCTTGAGTTGAGGCCTGGCCTGGGCGCTG
GGGCCGCCGCGTGCGAATCTGGTGGCACCTTCGCGCCTGTCTCGCTGCTTTCGATAAGTCTCTAGCCATTTAAA
ATTTTTGATGACCTGCTGCGACGCTTTTTTTCTGGCAAGATAGTCTTGTAAATGCGGGCCAAGATCTGCACACT
GGTATTTCGGTTTTTGGGGCCGCGGGCGGCGACGGGGCCCGTGCGTCCCAGCGCACATGTTCGGCGAGGCGG
GGCCTGCGAGCGCGGCCACCGAGAATCGGACGGGGGTAGTCTCAAGCTGGCCGGCCTGCTCTGGTGCCTGGC
CTCGCGCCGCCGTGTATCGCCCCGCCCTGGGCGGCAAGGCTGGCCCGGTCGGCACCAGTTGCGTGAGCGGAA
AGATGGCCGCTTCCCGGCCCTGCTGCAGGGAGCTCAAAATGGAGGACGCGGCGCTCGGGAGAGCGGGCGGGT
GAGTCACCCACACAAAGGAAAAGGGCCTTTCCGTCCTCAGCCGTCGCTTCATGTGACTCCACGGAGTACCGGG

CGCCGTCCAGGCACCTCGATTAGTTCTCGAGCTTTTGGAGTACGTCGTCTTTAGGTTGGGGGGAGGGGTTTTAT
GCGATGGAGTTTCCCCACACTGAGTGGGTGGAGACTGAAGTTAGGCCAGCTTGGCACTTGATGTAATTCTCCT
TGGAATTTGCCCTTTTTGAGTTTGGATCTTGGTTCATTCTCAAGCCTCAGACAGTGGTTCAAAGTTTTTTTCTTC
CATTTCAGGTGTCGTGACGTACGHHHHHHHHTCCAGTAGCAGACCTACGGCCACCACGCGGGGTTCTGCCGC
AGTGGATCGATGGGGATCCGAATTCGCCACCGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCAT
CCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCAC
CTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACC
ACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCG
CCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCG
AGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCA
ACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGA
ACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACC
AGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCT
GAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCT
CGGCATGGACGAGCTGTACAAGTAAGGCGCGCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTGGAATAAGG
CCGGTGTGCGTTTGTCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTG
GCCCTGTCTTCTTGACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAGGTCTGTTGAATGTC
GTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGTCTGTAGCGACCCTTTGCAGGCAGCGG
AACCCCCCACCTGGCGACAGGTGCCTCTGCGGCCAAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCA
CAACCCCAGTGCCACGTTGTGAGTTGGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACA
AGGGGCTGAAGGATGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCTTTAC
ATGTGTTTAGTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGTTTTCCTTTGAAAAACAC
GATGATAATATGGCCACAACCATGGTGAGCAAGGGCGAGGAGGATAACATGGCCATCATCAAGGAGTTCATG
CGCTTCAAGGTGCACATGGAGGGCTCCGTGAACGGCCACGAGTTCGAGATCGAGGGCGAGGGCGAGGGCCGC
CCCTACGAGGGCACCCAGACCGCCAAGCTGAAGGTGACCAAGGGTGGCCCCCTGCCCTTCGCCTGGGACATC
CTGTCCCCTCAGTTCATGTACGGCTCCAAGGCCTACGTGAAGCACCCCGCCGACATCCCCGACTACTTGAAGC
TGTCCTTCCCCGAGGGCTTCAAGTGGGAGCGCGTGATGAACTTCGAGGACGGCGGCGTGGTGACCGTGACCC
AGGACTCCTCCCTGCAGGACGGCGAGTTCATCTACAAGGTGAAGCTGCGCGGCACCAACTTCCCCTCCGACGG
CCCCGTAATGCAGAAGAAGACCATGGGCTGGGAGGCCTCCTCCGAGCGGATGTACCCCGAGGACGGCGCCCT
GAAGGGCGAGATCAAGCAGAGGCTGAAGCTGAAGGACGGCGGCCACTACGACGCTGAGGTCAAGACCACCT
ACAAGGCCAAGAAGCCCGTGCAGCTGCCCGGCGCCTACAACGTCAACATCAAGTTGGACATCACCTCCCACA
ACGAGGACTACACCATCGTGGAACAGTACGAACGCGCCGAGGGCCGCCACTCCACCGGCGGCATGGACGAGC
TGTACAAGTAA

5. Sequence 4.11 UGI encoding sequence and primers

UGI encoding sequence: NLS is highlighted in Yellow.

ATGACAAATCTGAGCGATATTATAGAAAAAGAGACTGGTAAACAGCTCGTGATTCAAGAGAGTATCCTTATG
CTGCCTGAGGAAGTGGAAGAAGTTATCGGCAATAAACCCGAGTCCGACATTCTGGTGCACACGGCGTATGAT
GAAAGCACCGACGAAAATGTGATGCTGCTTACTAGCGACGCTCCAGAGTACAAGCCATGGGCCCTGGTGATT
CAAGACAGTAACGGAGAGAATAAGATCAAAATGCTCTCCGGACTCAGATCTCGAGCTGATCCAAAAAAGAAG
AGAAAGGTAGATCCAAAAAAGAAGAGAAAGGTAGATCCAAAAAAGAAGAGAAAGGTA

BsiWI-UGI	TAGGGGCGTACGGCCACCATGACAAATCTGAGCGATATTATA
XhoI-UGI	TCAGCTCGAGATCTGAGTCCGGAGAGCATTTTGATCTTATTCTC

6. Sequence 4.12 ZF-AID-NLS/ZFP-AID$^{\Delta NES}$ sequences and primers

ZF-AID sequence: The NLS is highlighted in Yellow. ZF is in Red, the linker is in Green and the deaminase is in Blue. Of note, the nucleus export signal (NES) highlighted in Gray at the C-terminus of the deaminase was missing in ZF- **AID$^{\Delta NES}$**.

ATGGCTAGCCCCAGAGTGAGAACCGGTTCTAAGACACCTCCCCACGAGAGGCCTTTTCAGTGTAGAATTTGTA
TGCGTAATTTTTCTAGGTCCGATGTGCTGGCCAATCACACAAGGACTCACACTGGTGAAAAGCCCTTCCAATG
TAGAATTTGTATGCGCAATTTTTCTCAATCTTCTACTCTGACTAGACATCTGAGGACCCACACAGGCGAAAAG
CCTTTCCAGTGCAGAATTTGTATGAGAAATTTTTCTGAAAGACAGGGTCTGAAAAGACATCTGAAGACACATA

CAGGTGAAAAAGGATCCTCTGGTGGTGGTCTGGGTTCTACTGACAGCCTCTTGATGAACCGGAGGAAGTTTCT
TTACCAATTCAAAAATGTCCGCTGGGCTAAGGGTCGGCGTGAGACCTACCTGTGCTACGTAGTGAAGAGGCGT
GACAGTGCTACATCCTTTTCACTGGACTTTGGTTATCTTCGCAATAAGAACGGCTGCCACGTGGAATTGCTCTT
CCTCCGCTACATCTCGGACTGGGACCTAGACCCTGGCCGCTGCTACCGCGTCACCTGGTTCACCTCCTGGAGC
CCCTGCTACGACTGTGCCCGACATGTGGCCGACTTTCTGCGAGGGAACCCCAACCTCAGTCTGAGGATCTTCA
CCGCGCGCCTCTACTTCTGTGAGGACCGCAAGGCTGAGCCCGAGGGGCTGCGGCGGCTGCACCGCGCCGGGG
TGCAAATAGCCATCATGACCTTCAAAGATTATTTTTACTGCTGGAATACTTTTGTAGAAAACCACGAAAGAAC
TTTCAAAGCCTGGGAAGGGCTGCATGAAAATTCAGTTCGTCTCTCCAGACAGCTTCGGCGCATCCTTTTGCCCC
TGTATGAGGTTGATGACTTACGAGACGCATTTCGTACTTTGGGACTTTCCGGACTCAGATCTCGAGCTGATCCA
AAAAAGAAGAGAAAGGTAGATCCAAAAAAGAAGAGAAAGGTAGATCCAAAAAAGAAGAGAAAGGTA

BsiWI-ZF	ATAGGGGCGTACGGCCACCATGGCTAGCCCCAGAGTGAGAACCGGT
BsrGI-AID	TACTTGTACATTATACCTTTCTCTTCTTTTTTGGATCTACCTTTCTCTTCTTTTTTG GATCTACCTTTCTCTTCTTTTTTGGATCAGCTCGAGATCTGAGTCCGGAAAGTCC CAAAGTACGAAATGCGTCTCGTAA
BsrG1-ΔAID	CTTACTTGTACATTATACCTTTCTCTTCTTTTTTGGATCTACCTTTCTCTTCTTTTT TGGATCTACCTTTCTCTTCTTTTTTGGATCAGCTCGAGATCTGAGTCCGGACAGG GGCAAAAGGATGCGCCGAAG

7. Sequence 4.13 ZF_{GFPIN}- $AID^{\Delta NES}$ s/ZF_{GFPIN}Ns sequence and primers

The SV40 NLS is highlighted in Yellow. ZF_{GFPIN} (ZF_{GFPINL}/ZF_{GFPINR}) modules are in Red, nuclease/deaminase with the linkers is in Blue.

ZF_{GFPINL}- $AID^{\Delta NES}$

ATGGGACCTAAGAAAAAGAGGAAGGTGGCGGCCGCTGACTACAAGGATGACGACGATAAATCTAGACCCGG
GGAGCGCCCCTTCCAGTGTCGCATTTGCATGCGGAACTTTTCGCAGGACTCCTCCCTGCGGCGGCATACCCGT
ACTCATACCGGTGAAAAACCGTTTCAGTGTCGGATCTGTATGCGAAATTTCTCCCGGCAGGAGCACCTGGTGC
GGCATCTACGTACGCACACCGGCGAGAAGCCATTCCAATGCCGAATATGCATGCGCAACTTCAGTGACCCCA
CCTCCCTGAACCGGCATCTGAAGACACATACAGGTGAAAAAGGATCCTCTGGTGGTGGACTGGGGTCGACTG
ACAGCCTCTTGATGAACCGGAGGAAGTTTCTTTACCAATTCAAAAATGTCCGCTGGGCTAAGGGTCGGCGTGA
GACCTACCTGTGCTACGTAGTGAAGAGGCGTGACAGTGCTACATCCTTTTCACTGGACTTTGGTTATCTTCGCA
ATAAGAACGGCTGCCACGTGGAATTGCTCTTCCTCCGCTACATCTCGGACTGGGACCTAGACCCTGGCCGCTG
CTACCGCGTCACCTGGTTCACCTCCTGGAGCCCCTGCTACGACTGTGCCCGACATGTGGCCGACTTTCTGCGA
GGGAACCCCAACCTCAGTCTGAGGATCTTCACCGCGCGCCTCTACTTCTGTGAGGACCGCAAGGCTGAGCCCG
AGGGGCTGCGGCGGCTGCACCGCGCCGGGGTGCAAATAGCCATCATGACCTTCAAAGATTATTTTTACTGCTG
GAATACTTTTGTAGAAAACCACGAAAGAACTTTCAAAGCCTGGGAAGGGCTGCATGAAAATTCAGTTCGTCTC
TCCAGACAGCTTCGGCGCATCCTTTTGCCCCTGTATGAGGTTGAT

ZF_{GFPINR}- $AID^{\Delta NES}$

ATGGGACCTAAGAAAAAGAGGAAGGTGGCGGCCGCTGACTACAAGGATGACGACGATAAATCTAGACCCGG
GGAGCGCCCCTTCCAGTGTCGCATTTGCATGCGGAACTTTTCGTCCCAGACCCAGCTGGTGCGGCATACCCGT
ACTCATACCGGTGAAAAACCGTTTCAGTGTCGGATCTGTATGCGAAATTTCTCCCAGTCCACCACCCTGAAGC
GGCATCTACGTACGCACACCGGCGAGAAGCCATTCCAATGCCGAATATGCATGCGCAACTTCAGTCAGCGGA
ACAACCTGGGCCGGCATCTGAAGACACATACAGGTGAAAAAGGATCCTCTGGTGGTGGACTGGGGTCGACTG
ACAGCCTCTTGATGAACCGGAGGAAGTTTCTTTACCAATTCAAAAATGTCCGCTGGGCTAAGGGTCGGCGTGA
GACCTACCTGTGCTACGTAGTGAAGAGGCGTGACAGTGCTACATCCTTTTCACTGGACTTTGGTTATCTTCGCA
ATAAGAACGGCTGCCACGTGGAATTGCTCTTCCTCCGCTACATCTCGGACTGGGACCTAGACCCTGGCCGCTG
CTACCGCGTCACCTGGTTCACCTCCTGGAGCCCCTGCTACGACTGTGCCCGACATGTGGCCGACTTTCTGCGA
GGGAACCCCAACCTCAGTCTGAGGATCTTCACCGCGCGCCTCTACTTCTGTGAGGACCGCAAGGCTGAGCCCG
AGGGGCTGCGGCGGCTGCACCGCGCCGGGGTGCAAATAGCCATCATGACCTTCAAAGATTATTTTTACTGCTG
GAATACTTTTGTAGAAAACCACGAAAGAACTTTCAAAGCCTGGGAAGGGCTGCATGAAAATTCAGTTCGTCTC
TCCAGACAGCTTCGGCGCATCCTTTTGCCCCTGTATGAGGTTGAT

ZF_{GFPINL}-N

ATGGGACCTAAGAAAAAGAGGAAGGTGGCGGCCGCTGACTACAAGGATGACGACGATAAATCTAGACCCGG
GGAGCGCCCCTTCCAGTGTCGCATTTGCATGCGGAACTTTTCGCAGGACTCCTCCCTGCGGCGGCATACCCGT
ACTCATACCGGTGAAAAACCGTTTCAGTGTCGGATCTGTATGCGAAATTTCTCCCGGCAGGAGCACCTGGTGC
GGCATCTACGTACGCACACCGGCGAGAAGCCATTCCAATGCCGAATATGCATGCGCAACTTCAGTGACCCCA
CCTCCCTGAACCGGCACCTAAAAACCCACCTGAGGGGATCCCAACTAGTCAAAAGTGAACTGGAGGAGAAGA
AATCTGAACTTCGTCATAAATTGAAATATGTGCCTCATGAATATATTGAATTAATTGAAATTGCCAGAAATTC
CACTCAGGATAGAATTCTTGAAATGAAGGTAATGGAATTTTTTATGAAAGTTTATGGATATAGAGGTAAACAT
TTGGGTGGATCAAGGAAACCGGACGGAGCAATTTATACTGTCGGATCTCCTATTGATTACGGTGTGATCGTGG
ATACTAAAGCTTATAGCGGAGGTTATAATCTGCCAATTGGCCAAGCAGATGAAATGCAACGATATGTCGAAG
AAAATCAAACACGAAACAAACATATCAACCCTAATGAATGGTGGAAAGTCTATCCATCTTCTGTAACGGAATT
TAAGTTTTTATTTGTGAGTGGTCACTTTAAAGGAAACTACAAAGCTCAGCTTACACGATTAAATCATATCACTA
ATTGTAATGGAGCTGTTCTTAGTGTAGAAGAGCTTTTAATTGGTGGAGAAATGATTAAAGCCGGCACATTAAC
CTTAGAGGAAGTGAGACGGAAATTTAATAACGGCGAGATAAACTTT

ZF_{GFPINR}-N

ATGGGACCTAAGAAAAAGAGGAAGGTGGCGGCCGCTGACTACAAGGATGACGACGATAAATCTAGACCCGG
GGAGCGCCCCTTCCAGTGTCGCATTTGCATGCGGAACTTTTCGTCCCAGACCCAGCTGGTGCGGCATACCCGT
ACTCATACCGGTGAAAAACCGTTTCAGTGTCGGATCTGTATGCGAAATTTCTCCCAGTCCACCACCCTGAAGC
GGCATCTACGTACGCACACCGGCGAGAAGCCATTCCAATGCCGAATATGCATGCGCAACTTCAGTCAGCGGA
ACAACCTGGGCCGGCACCTAAAAACCCACCTGAGGGGATCCCAACTAGTCAAAAGTGAACTGGAGGAGAAG
AAATCTGAACTTCGTCATAAATTGAAATATGTGCCTCATGAATATATTGAATTAATTGAAATTGCCAGAAATT
CCACTCAGGATAGAATTCTTGAAATGAAGGTAATGGAATTTTTTATGAAAGTTTATGGATATAGAGGTAAACA
TTTGGGTGGATCAAGGAAACCGGACGGAGCAATTTATACTGTCGGATCTCCTATTGATTACGGTGTGATCGTG
GATACTAAAGCTTATAGCGGAGGTTATAATCTGCCAATTGGCCAAGCAGATGAAATGCAACGATATGTCGAA
GAAAATCAAACACGAAACAAACATATCAACCCTAATGAATGGTGGAAAGTCTATCCATCTTCTGTAACGGAA
TTTAAGTTTTTATTTGTGAGTGGTCACTTTAAAGGAAACTACAAAGCTCAGCTTACACGATTAAATCATATCAC
TAATTGTAATGGAGCTGTTCTTAGTGTAGAAGAGCTTTTAATTGGTGGAGAAATGATTAAAGCCGGCACATTA
ACCTTAGAGGAAGTGAGACGGAAATTTAATAACGGCGAGATAAACTTT

BsiWI-ZFL/R	TGGCAAATATTCTGAAATGAGCTGTTGACAATTAATCATCCGGTCCGTATAATCTGTG GAATTGTGAGCGGATAACAATTTCACACAAAGAGGAGAAAGGTACCATGTCGCGGA CCCGGCTCCC
BamHI-ZFL/R	AGAGGATCCTTTTTCACCTGTATGTGTCTTCAGATGCCGGCCCAGGTTGTTCCGCTGA C
NheI-ZFL/R	TGGCTAGCACCATGGGACCTAAGAAAAAGAGGAAGGTGGCGGCCGCTGACTACAAG GATGACGACGATAAATCTAGACCCGGGGAGCGCCCCTTCCAGTGTCG
ApaI-ZFL/R	CTTACCTTCGAAGGGCCCTTAATCAACCTCATAC

References

1. G. T. Group, H. Hospital, Therapeutic gene targeting, , 149–159 (1998).

2. E.-M. Händel, T. Cathomen, Zinc-finger nuclease based genome surgery: it's all about specificity., *Current gene therapy* **11**, 28–37 (2011).

3. S. G. Conticello, The AID/APOBEC family of nucleic acid mutators., *Genome biology* **9**, 229 (2008).

4. J. Boch *et al.*, Breaking the code of DNA binding specificity of TAL-type III effectors., *Science (New York, N.Y.)* **326**, 1509–12 (2009).

5. M. H. Porteus, Mammalian gene targeting with designed zinc finger nucleases., *Molecular therapy : the journal of the American Society of Gene Therapy* **13**, 438–46 (2006).

6. J. Zou *et al.*, Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells., *Cell stem cell* **5**, 97–110 (2009).

7. J. C. Miller *et al.*, A TALE nuclease architecture for efficient genome editing., *Nature biotechnology* **29**, 143–8 (2011).

8. M. Christian *et al.*, Targeting DNA double-strand breaks with TAL effector nucleases., *Genetics* **186**, 757–61 (2010).

9. C. Rada, J. M. Di Noia, M. S. Neuberger, Mismatch recognition and uracil excision provide complementary paths to both Ig switching and the A/T-focused phase of somatic mutation., *Molecular cell* **16**, 163–71 (2004).

10. S. M. D. Shandilya *et al.*, Crystal structure of the APOBEC3G catalytic domain reveals potential oligomerization interfaces., *Structure (London, England : 1993)* **18**, 28–38 (2010).

11. J. a Hurt, S. a Thibodeau, A. S. Hirsh, C. O. Pabo, J. K. Joung, Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based selection., *Proceedings of the National Academy of Sciences of the United States of America* **100**, 12271–6 (2003).

12. C. Prochnow, R. Bransteitter, M. G. Klein, M. F. Goodman, X. S. Chen, The APOBEC-2 crystal structure and functional implications for the deaminase AID., *Nature* **445**, 447–51 (2007).

13. H. H. Wang *et al.*, Programming cells by multiplex genome engineering and accelerated evolution., *Nature* **460**, 894–8 (2009).

14. F. Zhang *et al.*, LETTErs Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription, **29**, 149–154 (2011).

15. E.-M. Händel, S. Alwin, T. Cathomen, Expanding or restricting the target site repertoire of zinc-finger nucleases: the inter-domain linker as a major determinant of target site selectivity., *Molecular therapy : the journal of the American Society of Gene Therapy* **17**, 104–11 (2009).

16. M. Minczuk, M. a Papworth, J. C. Miller, M. P. Murphy, A. Klug, Development of a single-chain, quasi-dimeric zinc-finger nuclease for the selective degradation of mutated human mitochondrial DNA., *Nucleic acids research* **36**, 3926–38 (2008).

17. M. a Carpenter, E. Rajagurubandara, P. Wijesinghe, A. S. Bhagwat, Determinants of sequence-specificity within human AID and APOBEC3G., *DNA repair* **9**, 579–87 (2010).

18. L. Chelico, P. Pham, M. F. Goodman, Stochastic properties of processive cytidine DNA deaminases AID and APOBEC3G., *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **364**, 583–93 (2009).

19. J. Gorman, A. J. Plys, M.-L. Visnapuu, E. Alani, E. C. Greene, Visualizing one-dimensional diffusion of eukaryotic DNA repair factors along a chromatin lattice., *Nature structural & molecular biology* **17**, 932–8 (2010).

20. A.-M. Patenaude *et al.*, Active nuclear import and cytoplasmic retention of activation-induced deaminase., *Nature structural & molecular biology* **16**, 517–27 (2009).

21. H. Zan, P. Casali, AID- and Ung-dependent generation of staggered double-strand DNA breaks in immunoglobulin class switch DNA recombination: a post-cleavage role for AID., *Molecular immunology* **46**, 45–61 (2008).

22. C. D. Mol *et al.*, Crystal structure of human uracil-DNA glycosylase in complex with a protein inhibitor: protein mimicry of DNA., *Cell* **82**, 701–8 (1995).

23. I. a Klein *et al.*, Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes., *Cell* **147**, 95–106 (2011).

24. S. Ranjit *et al.*, AID binds cooperatively with UNG and Msh2-Msh6 to Ig switch regions dependent upon the AID C terminus., *Journal of immunology (Baltimore, Md. : 1950)* **187**, 2464–75 (2011).

25. K. Nishikura, Functions and regulation of RNA editing by ADAR deaminases., *Annual review of biochemistry* **79**, 321–49 (2010).

26. P. a Carr, G. M. Church, Genome engineering., *Nature biotechnology* **27**, 1151–62 (2009).

27. S. S. Brar *et al.*, Activation-induced deaminase, AID, is catalytically active as a monomer on single-stranded DNA., *DNA repair* **7**, 77–87 (2008).

28. J. C. Miller *et al.*, An improved zinc-finger nuclease architecture for highly specific genome editing., *Nature biotechnology* **25**, 778–85 (2007).

Chapter 5

Using engineered isogenic iPSC and heart-on-a-ChiP to modeling a mitochondrial cardiomyopathy

Gang Wang[1*], Megan L. McCain[2*],Luhan Yang[2,3], Aibin He[1], Francesco Silvio Pasqualini[2], Ashutosh Agarwal[2], Hongyan Yuan[2], Dawei Jiang[1], Lior Zangi[1], Judith Geva[1], Amy E. Roberts[1,4], Qing Ma[1], Jian Ding[1], Da-zhi Wang[1], Ronald J. A. Wanders[5], Wim Kulik[5], Frédéric M. Vaz[5], Michael A. Laflamme[6], Charles E. Murry[6,7], Kenneth R. Chien[8,], Richard I. Kelley[9], George M. Church[2,3],Kevin K. Parker[2¥], and William T. Pu[1,10,¥]

[1]Department of Cardiology, Boston Children's Hospital, Boston, MA, USA.
[2]Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, MA 02138, USA.
[3]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[4]BCH Department of Medicine, Division of Genetics, Boston Children's Hospital, Boston, MA, USA.
[5]Department of Clinical Chemistry and Pediatrics, Academic Medical Center, The Netherlands.
[6]Department of Pathology, Center for Cardiovascular Biology and Institute for Stem Cell and Regenerative Medicine, University of Washington, Seattle, WA 98109, USA.
[7]Departments of Bioengineering and Medicine, Seattle, WA 98109, USA.
[8]Department of Cell and Molecular Biology and Medicine, Karolinska Institutet, Stockholm, Sweden
[9]Division of Metabolism, Kennedy Krieger Institute, Baltimore, MD, USA.
[10]Harvard Stem Cell Institute, Harvard University, Cambridge, MA, USA.

Acknowledgements

This work was funded by support from the Barth Syndrome Foundation, the Boston Children's Hospital Translational Investigator Service, the NHLBI Progenitor Cell Biology Consortium (NIH U01 HL100401 and U01 HL100408), NIH RC1 HL099618, NIH UH2 TR000522, NIH NHRGI CEGS grant P50 HG005550, and charitable donations from Edward Marram, Karen Carpenter, and Gail Federici Smith.

Author contribution

G.W and W.T.P conceived the study jointly with M.L.M., L.Y. designed and performed the experiments with assistance from other co-authors.; G.W and W.T.P wrote the manuscripts with the help of all the other co-authors. G.M.C. supervised L.Y.. In particular, I generated the isogeneic stem cell lines and joined the functional studies with Gang Wang.

Adopted from submitted Manuscript "Modeling a mitochondrial cardiomyopathy and its correction using iPSC and heart-on-chip technologies"

Summary

Patient specific iPSCs promise to offer new insights into human disease pathogenesis. However, significant hurdles remain in developing disease models that replicate disease pathophysiology. Here, we demonstrated that we can use the combination of patient derived hiPSCs and genetically engineered iPSCs to model the mitochondrial cardiomyopathy of Barth syndrome (BTHS). Using Cas9, we generated isogenic iPSCs carrying the mutation identified in BTHS patients. CM derived from patients' and engineered hiPSC recapitulated characteristic disease metabolic abnormalities. BTHS iPSC-CMs did not assemble myofibers normally, and engineered BTHS "heart on chip" tissues contracted poorly. Replacement of the defective gene product, or supplementation with a precursor to a depleted metabolite, corrected metabolic abnormalities, myofibrillogenesis, and contractile activity of BTHS iPSC-CMs. These data indicate that the combination of patient-specific iPSCs and genetically engineered iPSCs with organ-on-chip models form an effective platform to reveal the cellular etiology of disease and to identify potential therapeutic strategies.

Introduction

Research into the pathogenesis of cardiomyopathy has historically been hindered by the lack of suitable model systems. Cardiomyocyte differentiation of patient-derived induced pluripotent stem cells (iPSCs) offers one promising avenue to surmount this barrier, and reports of iPSC modeling of cardiomyopathy have begun to emerge (*1–3*). However, realization of this promise will require approaches to overcome genetic heterogeneity of patient-derived iPSC lines and to assay contractile function of tissue constructs assembled from iPSC-derived

cardiomyocytes (iPSC-CMs). Here, we combined genome-edited iPSCs, modified RNA (modRNA) (*4*), and "heart on a chip" (*5*) technologies to replicate the pathophysiology of Barth syndrome cardiomyopathy in tissue constructs. Furthermore, we use the bioengineered tissue constructs to model the genetic and metabolite-induced correction of the Barth disease phenotype.

Results

An iPSC-CM model of Barth syndrome

Barth syndrome is an X-linked cardiac and skeletal mitochondrial myopathy caused by mutation of the gene Tafazzin (*TAZ*) (*6*), an acyltransferase responsible for normal acylation of cardiolipin (CL), the major phospholipid of the mitochondrial inner membrane (*7*). The pathogenesis of cardiomyopathy in Barth syndrome is poorly understood. We generated iPSCs from two unrelated individuals with Barth Syndrome under institutionally approved protocols (Supplemental Materials). Two lines, BTH-H and BTH-C, reprogrammed using retroviral or modified RNA approaches (*8*), respectively. As controls, we used three normal iPSC lines, generated by retroviral (2) or modified RNA (1) reprogramming. We next differentiated the iPSCs into iPSC-derived cardiomyocytes (iPSC-CMs) using an established protocol (*9*)

Having established the cell lines, we investigated the phenotype of BTHS iPSC-CMs. BTHS is characterized by depletion of mature CL and accumulation of an immature form, monolysocardiolipin (MLCL) (*10*, *11*). This hallmark of BTHS was recapitulated in the patient-derived iPSC-CMs, in which phospholipid mass spectrometry showed that the ratio of MLCL to CL in BTHS iPSC-CMs exceeded 0.3, the clinically used diagnostic threshold (*11*) (Figure 5_1

A-B). To determine if energy metabolism was perturbed in BTHS iPSC-CMs, we measured cellular ATP levels in iPSC-CMs cultured in galactose, which does not support glycolysis. BTHS iPSC-CM ATP levels were significantly lower than controls (Figure 5_1 C). Consistent with cellular energy deprivation, AMP-dependent kinase (AMPK) was markedly activated in BTHS iPSC-CMs, as demonstrated by immunoblotting with activation-state-specific antibodies (Figure 5_1 D).

We went further to investigate BTHS iPSC-CM metabolic activity. Despite lower basal ATP levels, BTHS iPSC-CMs had a higher basal and F1F0 ATP synthase-dependent oxygen consumption rate (OCR) than control iPSC-CMs (Figure 5_1 E-G), indicative of inefficient mitochondrial ATP generation. Oligomycin-independent oxygen consumption ("H^+ leak") was also increased in BTHS iPSC-CMs (Figure 5_1H), while maximal electron transport chain activity ("Respiratory Capacity") was severely impaired (Figure 5_1I). As a result of increased basal and depressed maximal OCR, BTHS iPSC-CMs had markedly decreased respiratory reserve (Figure 5_1 J).

Decreased ATP levels in the setting of increased F1F0 ATP synthase-dependent oxygen consumption suggested that CL abnormalities impaired F1F0 ATP synthase activity. We tested this hypothesis by measuring F1F0 ATP synthase specific activity by selective complex immunocapture followed by assays of complex quantity and activity. While expression was comparable between BTHS and control iPSC-CMs, activity was lower in BTHS iPSC-CMs (Figure 5_1K). Collectively, the data demonstrate that TAZ deficiency and consequent CL

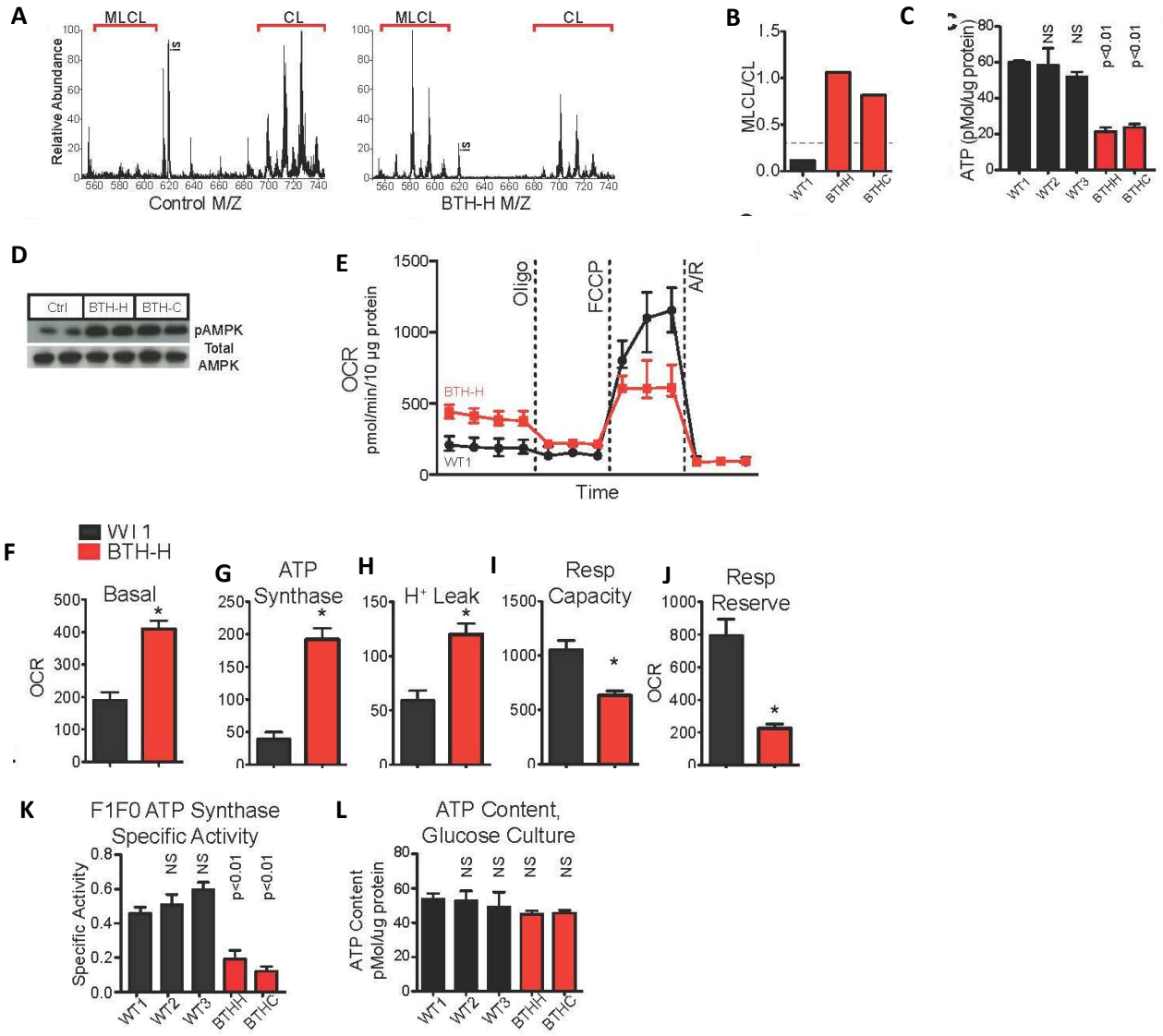


Figure 5_1. Mitochondrial abnormalities in BTHS iCMs.
(A) Mass spectrum of control and BTH-H CL, showing depletion of mature CL and accumulation of its immature form (MLCL) in BTHS iPSC-CMs.
(B) Comparison of MLCL/CL ratio in BTH-H, BTH-C, and control iPSC-CMs. The dashed line indicates the clinical
diagnostic threshold for BTHS.
(C) ATP levels in BTHS and control iPSCCMs cultured in galactose. n=3.
(D) AMPK activation in BTHS iPSC-CMs cultured in galactose. Activated and total AMPK were measured by quantitative western blotting.
(E-G) Abnormal BTH-H iPSC-CM mitochondrial function. Function was measured using cellular oxygen consumption rate (OCR), normalized to total protein. Oligo, oligomycin. FCCP, carbonyl cyanide-4-(trifluoromethoxy)phenylhydrazone. A/R, antimycin plus rotenone. Measures of mitochondrial function (defined in Fig. S4) were quantitatively compared between control and BTH-H iCMs. n=3. *, P<0.05.
(K) Measurement of F1F0 ATP synthase specific activity. Total activity was normalized to the amount of F1F0 ATP synthase, measured by ELISA. n=3.
(L) BTHS iPSC-CM and control ATP levels were comparable when cultured in glucose. n=3.

abnormalities decrease basal mitochondrial ATP generating efficiency by impairing F1F0 ATP synthase activity and by reducing peak mitochondrial electron transport chain function.

Measurement of cardiac energy stores by nuclear magnetic resonance spectroscopy in a Barth syndrome patient with cardiomyopathy indicated that myopathy can occur without measurable energy depletion in at least some BTHS patients, suggesting that recruitment of compensatory metabolic pathways can normalize cardiac energy levels. We tested this hypothesis by culturing BTHS iPSC-CMs in glucose, which supports both glycolysis and oxidative phosphorylation. Interestingly, glucose normalized BTHS iCM ATP levels and reversed the elevated basal oxygen consumption rate (Figure 5_1L). These observations indicate that glucose restores ATP levels through alternative metabolic pathways, thereby reducing basal F1F0 ATPase activity, but does not correct underlying mitochondrial defects in the electron transport chain.

Using engineered isogenic CM to test whether TAZ mutation causes BTHS phenotypes

Because BTHS patient-derived and control iPSCs had numerous genetic differences other than TAZ mutation, we took two independent approaches to further establish the causative role of TAZ mutation in abnormal CL biogenesis and mitochondrial function.

First, we sought to test whether TAZ mutation is sufficient for the phenotype by re-introducing WT TAZ into BTHS iPSC-CMs. To this end, we synthesized TAZ mRNA, substituting 5-methylcytidine for cytidine, and pseudouridine for uridine. This modified RNA ("modRNA") has minimal toxicity (*4*, *8*) and efficiently transfects cardiomyocytes (Figure 5_2A). TAZ modRNA likewise transfected iPSC-CMs and TAZ protein localized to mitochondrial function in BTHS iPSC-CMs, although maximal respiratory capacity was rescued

incompletely (Figure 5_2 B-E). These results indicate that TAZ restoration rapidly corrects the BTHS mitochondrial phenotype.

Second, we used Cas9-mediated scarless genome editing (*12*) to mutate TAZ in the control human line PGP1-iPSC, yielding three iPSC lines that are isogenic except for the sequence at TAZ exon 6 (Figure 5_3 A-D). PGP1-BTHH contains the TAZ frameshift mutation from the BTH-H line (1 nt deletion), while PGP1-NHEJ contains a distinct frameshift mutation at the same site (14 nt insertion). PGP1-NT is a control line handled in parallel to PGP1-BTHH and PGP1-NHEJ, but without a TAZ mutation. We verified the pluripotency of the cell line (Figure 5_3 E-G) and differentiated into iPSC-CMs. iPSC-CMs derived from these isogenic TAZ mutant lines fully recapitulated the cardiolipin, mitochondrial, and ATP deficits that we observed in patient-derived iPSCs and in the neonatal rat TAZ knockdown model (Figure 5_4 A-D). Together, these data indicate that TAZ mutation alone is sufficient to cause these phenotypes in a control genetic background.

Abnormal sarcomerogenesis in BTHS iPSC-CMs

Mitochondria regulate cardiomyocyte maturation (*13*), a hallmark of which is assembly of organized arrays of sarcomeres. To test whether TAZ deficiency causes defective sarcomeres arrangmetn, we engineered iPSC-CM shape by seeding the cells on micropatterned fibronectin rectangles designed to mimic the dimensions of human adult cardiomyocytes (*14*), with length:width ratios of approximately 7:1 (95 μm X 13 μm). While sarcomeres in control iPSC-CMs extended serially across the entire length of the cell, sarcomeres in patient-derived BTH-H iPSC-CMs were intermittent and sparse (Figure 5_5 A). In BTH-H iPSC-CMs, sarcomeric

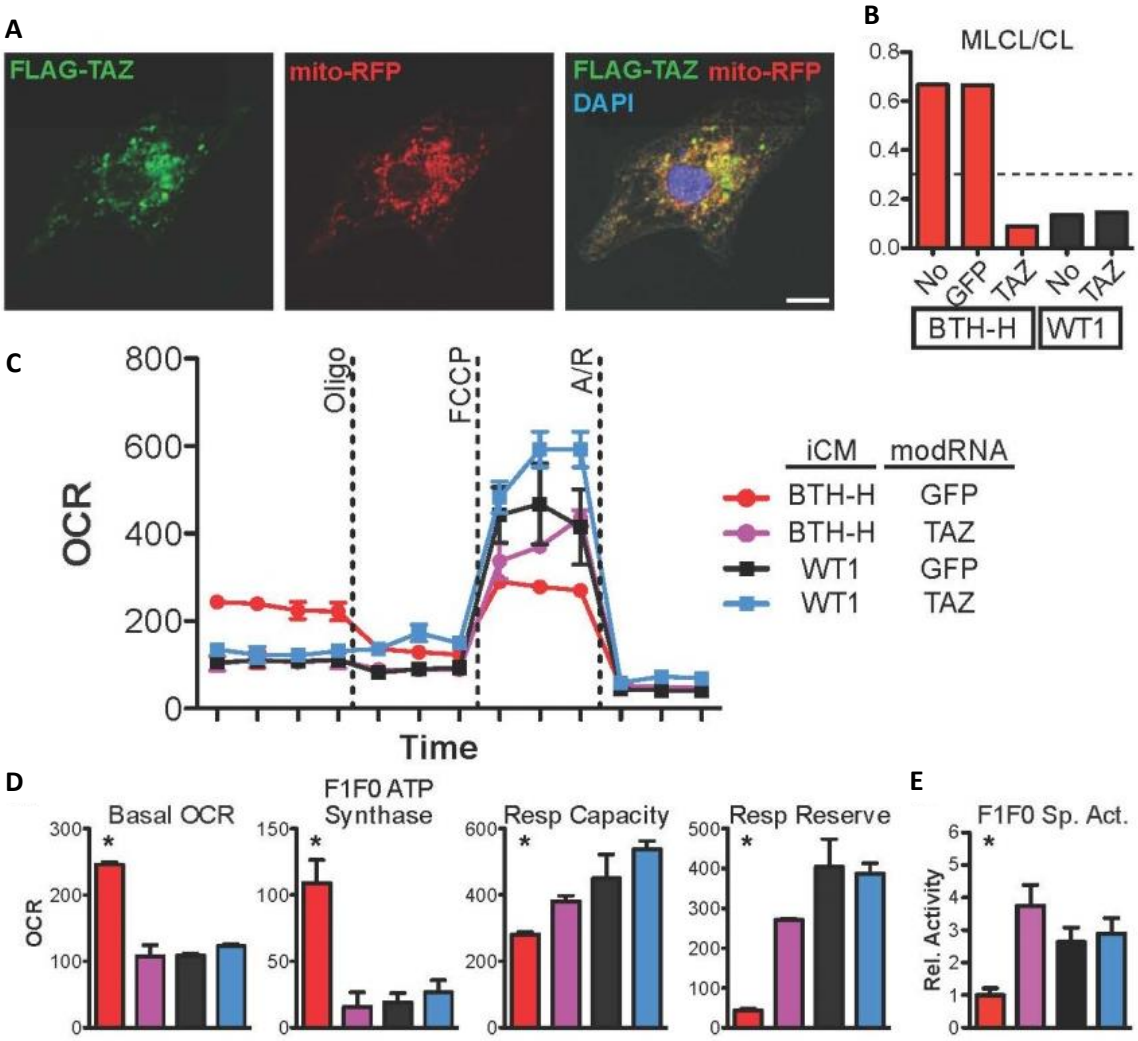


Figure 5_2. TAZ deficiency is necessary and sufficient to cause the iPSC-CM metabolic phenotype.
(A) ModRNA encoding FLAG-tagged TAZ was transfected into iCMs. Mitochondria were labeled with virally delivered RFP with a mitochondrial localization sequence. FLAG co-localized with RFP. Bar = 10 μm.
(B) TAZ modRNA restored cardiolipin biogenesis. BTH-H or control iCMs were transfected with the indicated modRNA and cardiolipin composition was measured by mass spectroscopy.
(C) Mitochondrial function testing showed that TAZ modRNA normalized mitochondrial function in BTHS iCMs.
(D)Quantitation of mitochondrial functional parameters from c. n=3.
(E) F1F0 ATP Synthase specific activity. n=6. *, P<0.05 compared to each other group

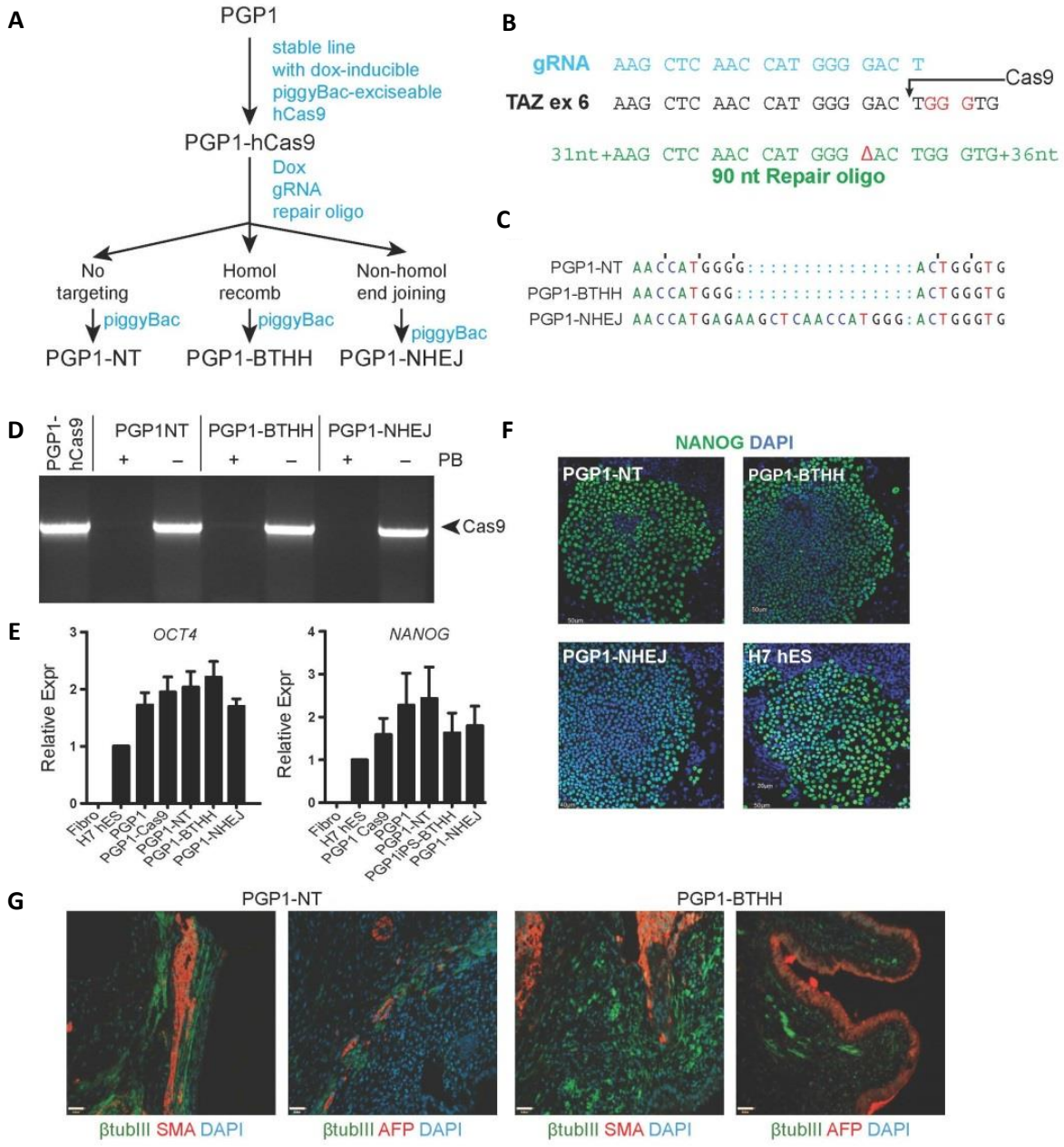


Figure 5_3 Generation and characterization of isogentic iPSCs through Cas9-mediated genome editing
(A) Schematic of modified genome editing approach
(B) Targeting strategy
(C) Sequences of targeted region of exon 6 in clones used in the study
(D) PiggyBac-mediated removal of Cas9 expression cassette. Clones were genotyped using Cas9 primers
(E) Expression of the pluripotency marker in genome-edited clones
(F) Expression of the pluripotency marker Nanog in genome-edited clones
(G) Formation of tissues from all three germ layers in teratoma assays from two of the clones used in this study

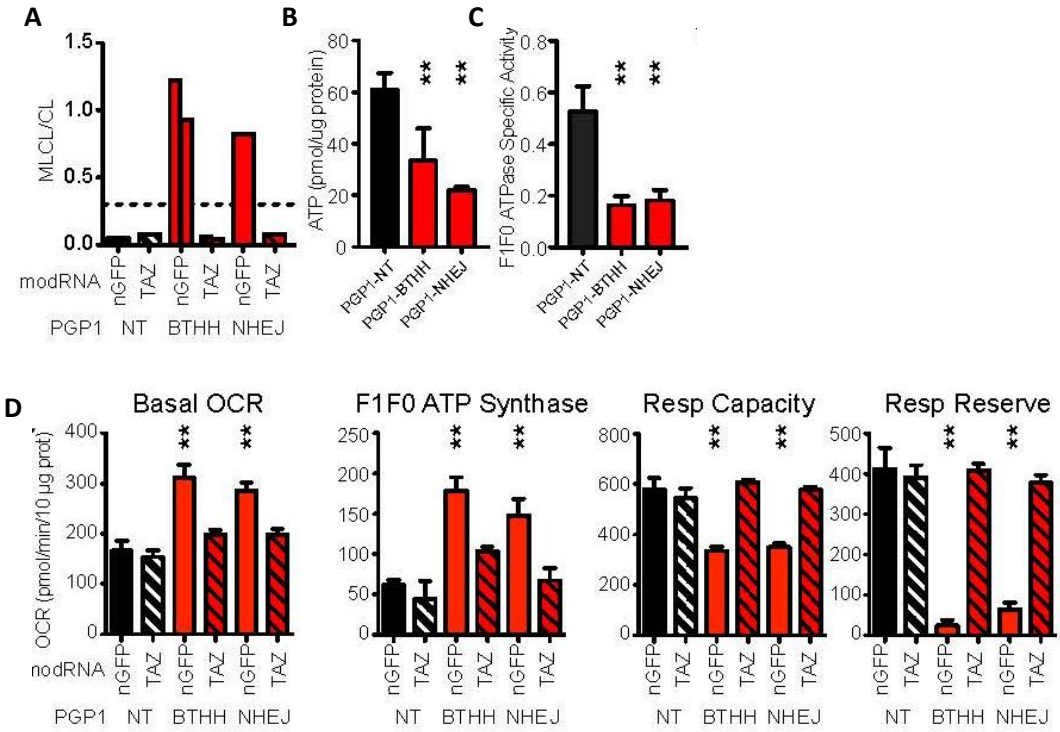


Figure 5_4 Functional analysis of isogentic iPSCs
(A) Cardiolipin maturation abnormalities in CM dereived from PGP1 carrrying TAZ mutation but not isogenic control iPSC-CMs
(B) Decreased basal ATP level of TAZ mutant compared to isogenic control iPSC-CM culture in galactose media
(C) Decrased F1F0 ATPase specificity activity in TAZ mutant compared to isogenic control iPSC-CMs
(D) Abnormalities of mitochondrial function in TAZ mutant iPSC-CMs. Abnormalities were rescued by TAZ modRNA transfection.

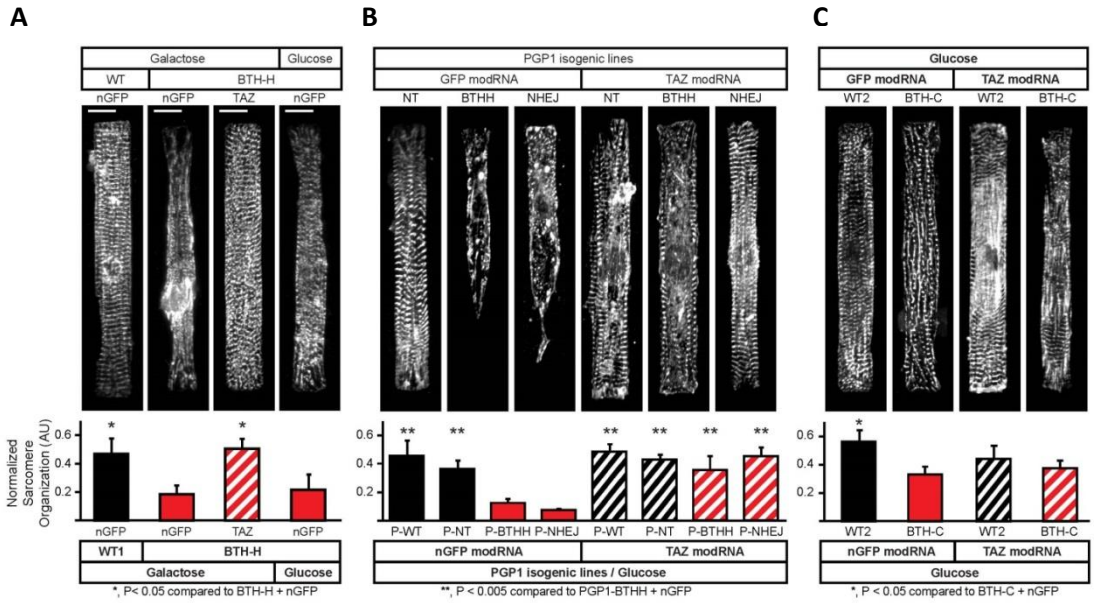


Figure 5_5 Reduced sarcomere organization and contractile function in BTHS iPSC-CMs
(A) Representative images of a-actinin-stained iPSC-CMs cultured on micropatterned fibronectin rectangles with length:width ratios of 7:1. an unbiased metric of sacromeric organization, showed significantly impared myofibrillar assembly in BTHS iCMs, which was rescued by TAZ modRNA but not glucose culture. Bar= 10 m
(B) Representative images of a-actinin-stained isogenetic hiPSCs cultured on micropatterned fibronectin rectangles with length:width ratios of 7:1. an unbiased metric of sacromeric organization, showed significantly impared myofibrillar assembly in BTHS iCMs, which was rescued by TAZ modRNA but not glucose culture. Bar= 10 m
(C) iPSC-CM engineered tissue was cultured on micropatterned muscular thin film (MTF) substrate. Cardiomyocyte force generation reduces the radius of curvature of the construct while contracting from diastole to peak systole.

organization was lower compared to control and was recovered to by TAZ modRNA treatment (Figure 5_5 A). Although glucose culture normalized ATP levels to a comparable degree compared to TAZ modRNA, glucose culture did not normalize sarcomere formation (Figure 5_5 A). The defects in sarcomere assembly were recapitulated in genome edited PGP1-BTHH and PGP1-NHEJ iPSC-CMs compared to isogenic PGP1-NT controls (Figure 5_5B), confirming the causative role of the BTH-H exon 6 frameshift TAZ mutation. These data suggest that myofibrillogenesis and cell size are sensitive to mitochondrial function independent of whole cell ATP levels. Interestingly, BTH-C iPSC-CMs exhibited sarcomere organization that was not significantly different from controls (Figure 5_5 C). This phenotypic heterogeneity may be due to the specific BTH-C missense mutation, or it may be due other variables between cell lines. Further use of genome-edited cell lines will be necessary to better understand genotype-phenotype relationships in this disease.

Myocardial constructs model BTHS and its genetic rescue

We asked if we could replicate the pathophysiology of Barth Syndrome in an in vitro model of engineered myocardium and demonstrate the efficacy of the TAZ modRNA treatment on the disease. We used our "heart on a chip" assay (*5*) to quantitatively measure contractility of myocardial tissue assembled from BTHS or control iPSC-CMs. MACS-selected iPSC-CMs were seeded onto thin elastomers supported by glass coverslips (*15*, *16*). Over a five day culture period, the iPSC-CMs self-organized into laminar, anisotropic myocardium. Pre-cut muscular thin films (MTF) were then peeled from the glass substrate, allowing them to contract away from the plane of the coverslip (Figure 5_6 A). From the radius of curvature of each MTF measured

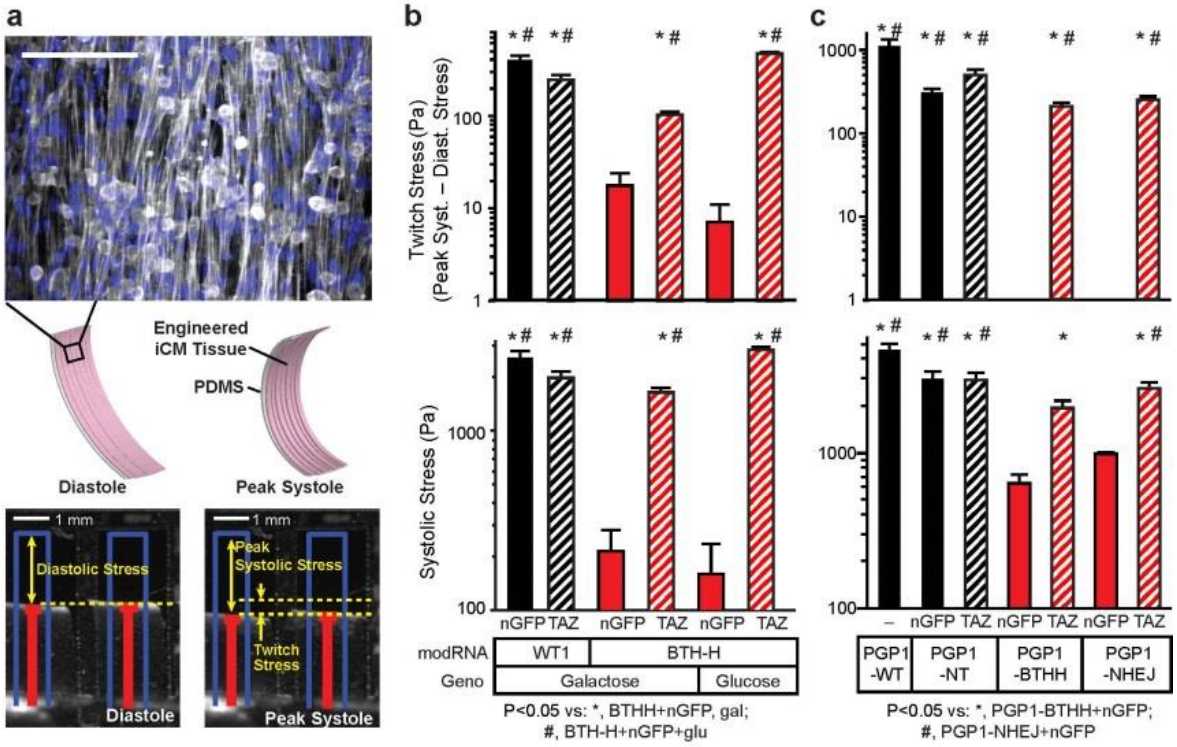


Figure 5_6 Depressed force generation by BTHS myocardial tissue constructs
(A) iPSC-CMs seeded onto thin elastomers with linearly patterned fibronectin formed selforganizing anisotropic myocardial tissues. Cardiomyocyte force generation reduces the radius of curvature of the construct while contracting from diastole to peak systole. Red lines indicate automated tracking of muscular thin film (MTF) projected onto the horizontal plane.
(B) and (C) Twitch stress and peak systolic stress generated by MTFs from patient-derived BTH-H and control iPSC-CMs (b), genome-edited TAZ frameshift and control iPSC-CMs (c), Statistical comparisons by Kruskal-Wallis One Way ANOVA on ranks and Dunn's post-hoc test. Sample size is indicated by number inside each bar.

throughout the myocardial tissue contraction cycle, we calculated the diastolic and peak systolic stresses using a modification of Stoney's equation (*5*, *15*, *17*). Patient-derived BTH-H iPSC-CM tissues in galactose were significantly weaker (Figure 5_6 B) over the same stimulation frequency range, indicating that BTHS engineered myocardial tissue recapitulates the BTHS myopathic phenotype.

Next we asked if the engineered myocardial tissue constructs effectively model diseae correction. We reintroduced TAZ by treating with modRNA for 5 days, then measured myocardial tissue construct function. Treatment of BTH-H iPSC-CMs (Figure 5_6 B) with TAZ modRNA restored contractile function to levels comparable to controls, further confirming that this phenotype was revsible and due to TAZ mutation.

We assessed the contribution of whole cell ATP to contractile dysfunction by culturing engineered iPSC-CM tissues on MTF substrates in glucose, which increased cellular ATP levels to a level comparable to TAZ modRNA. In contrast to TAZ modRNA rescue, glucose culture alone did not restore BTH-H iPSC-CM force generation at any tested stimulation frequency (Figure 5_6 B). Together, these data show that BTH-H iPSC-CMs have a severe defect in contractility that occurred regardless of the energetic substrate.

To further confirm these results, we generated BTHS and isogenic control myocardial tissue constructs using the genome-edited iPSCs. Contractile function of the mock-manipulated PGP1-NT iPSC-CM myocardial constructs was equivalent to the parental PGP1 cell line, while TAZ disruption in both PGP1-BTHH and PGP1-NHEJ caused severe loss of force generation (Figure 5_6 C). Reintroduction of TAZ by modRNA restored contractile force production (Figure 5_6C).

Together, these results indicate that TAZ mutation is sufficient to cause a myopathic phenotype in myocardial tissue constructs, and that this phenotype is readily reversible upon TAZ replacement.

Discussion

One obstacle to wider use of iPSC disease models has been the genetic and epigenetic variation between cell lines, which introduces confounding variables that can be difficult to control. We show that Cas9-mediated genome editing is an excellent strategy to isolate a mutation of interest and show that it is sufficent to cause a disease phenotype. Gene replacement using modified RNA technology is another highly portable approach that demonstrates the acute requirement of a gene mutation for a disease phenotype within a given cell line.

Our metabolic and functional analysis of human Barth syndrome cardiomyocytes elucidated mitochondrial functional impairment caused by mature CL depletion. Our data show that the contracile deficit of BTHS cardiomyocytes is not a result of global cellular energy depletion. Although subtle defects in local ATP concentrations cannot be excluded, these data support the notion that cardiomyopathy in BTHS results from an ATP-independent role of mitochondria in sarcomere assembly and contractile activity. The TAZ frameshift mutation hindered both sarcomere assembly, and likely this contributed to impaired systolic and twitch stress in tissue constructs assembled from these iPSC-CMs. On the other hand, the TAZ point mutation that we studied severely impaired twitch stress but not sarcomere assembly or systolic stress, indicating additional mechanisms linking TAZ mutation to impaired contracile force generation. These observations have implications for more common diseases such as ischemic and diabetic cardiomyopathy, since mature cardiolipin depletion also occurs in these conditions.

In conclusion, we show that the combination of tissue engineering, induced pluripotency, and genome editing models human heart disease and its genetic and pharmacological correction. Prior studies on diseased human iPSC-CMs characterized the properties of individual cells, but cardiomyocytes function as components of a highly integrated tissue. In addition, single cell contraction assays are complicated by substantial cell-to-cell variation. Using tissue engineering approaches, we built control and diseased muscular thin film myocardial tissue constructs and showed that these tissues model the cardiomyopathic phenotype seen in patients. Furthermore, we envision that the disease-specific "heart-on-chip" assay would be useful for preclinical assessment of candidate therapies.

Materials and Methods

Cell culture

Cell lines used in this study are summarized in Table 5_1. Low passage skin fibroblasts were obtained from skin biopsies from two unrelated BTHS patients with informed consent under a Boston Children's Hospital IRB approved protocol. Control cells were derived from BJ cells (Stemgent). The BTH-H iPSC line was established by retroviral delivery of three reprogramming factors (SOX2, KLF4 and OCT4), followed by application of the histone deacetylase inhibitor valproic acid (VPA) as described[1]. The BTH-C iPSC line was established by modRNA reprogramming as described[2]. Several iPSC clones with ES cell morphology and with positive vital staining for TRA-1-81 or TRA-1-60 staining[3], were further characterized to yield the final two lines studied. Karyotyping was performed by Cell Line Genetics, Inc. Teratomas were formed by injection of 10^6 iPSCs intramuscularly into the flanks of adult SCID mice. Teratomas were examined by H&E staining.

Table 5. 1. Human iPSCs used in this study

Name	Starting Cells	Derivation method	Lab source	TAZ mutation	Genome editing	Reference
WT1iPS	Male normal newborn Fibroblasts	Retrovirus OSKM	Pu lab			
WT2iPS	Male normal adult fibroblasts	Retrovirus OSKM	Daley lab			PMID: 18691744
WT3iPS	Male normal adult fibroblasts	Modified mRNA OSKML	Allele Biotech			PMID: 22984641
BTHHiPS	Male adult patient fibroblasts	Retrovirus OSKM	Pu Lab	Exon6 GAC>ACT delete G		
BTHCiPS	Male adult patient fibroblasts	Modified mRNA OSKML	Pulab	Exon4 TCC>CCC		
PGP1iPS	Male normal adult fibroblasts	Retrovirus OSKM	Church lab/Coreill Institute GM23338		Dox inducible Cas9 expression	
PGP1iPS NT (No targeting)	Male adult fibroblasts	Retrovirus OSKM	Church lab/Pulab		Cas9 induced without TAZ mutation	
PGP1iPS BTHH	Male adult fibroblasts	Retrovirus OSKM	Church lab/PuLab	Exon6 GAC>ACT delete G	Homologous recombination introduction of BTHH mutation	
PGP1iPS NHEJ	Male adult fibroblasts	Retrovirus OSKM	Church lab/Pulab	TAZ exon 6 insertion	NHEJ-induced TAZ mutation	

Cardiomyocyte differentiation was induced as previously reported[4], with minor modifications. Cells were detached by 3-5 min incubation with Versene (Invitrogen) and seeded onto Matrigel-coated plates at a density of 10,000 cells/cm^2 in murine embryonic fibroblast conditioned medium (MEF-CM) plus 4 ng/mL bFGF for 2-3 days before induction. Cells were covered with matrigel (1:60 dilution) on the day before induction. To induce cardiac differentiation, we replaced MEF-CM with RPMI+B27 medium (RPMI-1640, 2 mM L-glutamine, x1 B27 supplement without insulin) supplemented with 100 ng/mL of Activin A (R&D Systems) for 24 hours, followed by 10 ng/mL human bone morphogenetic protein 4 (R&D) and 10 ng/mL human basic fibroblast growth factor for 4 days without culture medium changes. The culture medium was subsequently replaced with RPMI+B27 supplemented with 100 ng/mL of DKK1 (R&D) for 2 days. At day 7, the culture medium was changed to RPMI+B27 without supplementary cytokines; culture medium was refreshed every 1-2 days. Leibovitz L-15 medium was substituted for RPMI for galactose containing culture media.

NRVMs were isolated from neonatal rat heart ventricles by collagenase digestion using the Neomyts isolation kit (Cellutron). Procedures involving animals were performed under protocols approved by the Institutional Animal Care and Use Committee.

Mitochondria were labeled using BacMam 2.0 mitochondria-RFP (Invitrogen), in which baculovirus delivers RFP tagged with the mitochondrial localization sequence of E1 alpha pyruvate dehydrogenase.

Cas9 Genome Editing

In brief, we constructed a piggyBac transposon that expresses the reverse tet activator and a human codon optimized Cas9 under the control of a tet response element. Transient transfection of PGP1-iPSCs (Coriell) with piggyBac and this engineered transposon yielded PGP1iPSC-hCas9. We subsequently designed guide RNA and donor oligonucleotides to introduce the BTHH TAZ mutation into exon 6. After transient dox administration and transfection with gRNA and donor oligonucleotides, we screened individual clones by Sanger sequencing. We selected an unmodified clone, a clone containing the BTHH mutation (homologous recombination), and a clone containing a novel insertion due to non-homologous end joining. Transient piggyBac transfection subsequently led to removal of the Cas9-containing transposon.

Cardiac ^{31}P-NMR

Cardiac ^{31}P-NMR was performed under institutionally approved protocols as described[5]. High energy phosphate stores were estimated from the phosphocreatine peak area, normalized to the peak area from the β-phosphate of ATP[5].

iPSC-CM characterization

For characterization, iPSC-CMs were dissociated with Accumax (Innovative Cell Technologies) on day 11-12 of differentiation. Dissociated cells were stained with anti-VCAM1 antibody (refer to Table 5_1 for antibody information) conjugated with allophycocyanin (APC) and magnetically sorted using anti-APC microbeads (Miltenyi Biotec). For validation of MACS,

sorted cells were fixed and stained with TNNT2-Alexa-488. Data were analyzed with DIVA (BD) and FlowJo (Treestar) software.

For mitochondrial function assays, 60,000 sorted cells were seeded in 0.1% gelatin-coated seahorse assay wells in alpha MEM (Invitrogen) with 10% FBS. They were then changed to L-15 media supplemented with 1x B27 supplement without insulin for 5-7 days. After measurement of oxygen consumption rate, total protein levels of iPSC-CMs was determined using the BCA protein assay (Thermo Scientific). OCR was measured using a Seahorses Biosciences XF24 extracellular flux analyzes and normalized to total protein. For ATP assays, cells were grown in indicated media and supplements for 5 days. ATP assay reagent (Promega) was added directly to wells and light output was measured with a plate luminometer. Readout was normalized to total protein.

Electron microscopy was performed on a Tecnai G Spirit BioTWIN instrument. At least 10 randomly selected fields containing cardiomyocytes were imaged per sample.

Quantitation of sarcomere assembly for iPSC-CMs plated on anisotropic gelatin substrate was performed on randomly acquired confocal images. Images were overlaid with a point-counting grid, which was used to measure overall cell area and cell area containing linearly organized ACTN2 staining.

Gene expression analysis

Samples were fixed in 4% paraformaldehyde and permeabilized with 0.05% Triton X-100. Imaging was performed on an Olympus FV1000 or Zeiss LSM 5 LIVE confocal microscope. qRTPCR was performed from total RNA using primers listed in Table 5_2. Sybr green chemistry was used for real time PCR detection on an ABI 7500 instrument.

Modified RNA synthesis and delivery

The OCT4, SOX2, KLF4, cMYC, and LIN28 cDNA templates were obtained from Addgene (plasmids 26815-9). The TAZ modRNA cDNA template was expressed from pcDNA3.3-TOPO-T7-5'UTR-cMyc-3'UTR, which contains the T7 promoter and optimized 5' and 3' untranslated regions (Addene plasmid 26818). The vector was modified to place unique AscI and NheI restriction sites between the 5' and 3' untranslated regions. Full length human TAZ cDNA corresponding to Refseq NM_000116 with a 5' FLAG tag was cloned into modified vector.

To synthesize the modRNA, the UTRs and ORF were PCR amplified using a polyA-tailed primer. 1.6 µg of purified PCR productwas transcribed in a 40 µl reaction system using the MEGAscript T7 kit (Ambion, Austin, TX) and a custom ribonucleoside cocktail from Allele Biotechnology (No. ABP-PP-NTPMIX), containing pseudouridine-5´-triphosphate, Methylcytidine-5´-triphosphate, GTP, ATP, and ARCA (Cap Analog). Reactions were incubated

Table 5.2. Oligonucleotide Primers used in this study.

Name	Sequence
OCT-4F	AGTTTGTGCCAGGGTTTTTG
OCT-4R	ACTTCACCTTCCCTCCAACC
NANOGF	TTTGGAAGCTGCTGGGGAAG
NANOGR	GATGGGAGGAGGGGAGAGGA
NKX2.5F	CATTTACCCGGGAGCCTACG
NKX2.5R	GCTTTCCGTCGCCGCCGTGCGCGTG
MYL7F	GAGGAGAATGGCCAGCAGGAA
MYL7R	GCGAACATCTGCTCCACCTCA
BrachyuryF	CGGAACAATTCTCCAACCTATT
BrachyuryR	GTACTGGCTGTCCACGATGTCT
Vcam1F	CCGGATTGCTGCTCAGATTGGA
Vcam1R	AGCGTGGAATTGGTCCCCTCA
Human GAPDHF	GTGGACCTGACCTGCCGTCT
Human GAPDHR	GGAGGAGTGGGTGTCGCTGT
Rat TAZF	TGAAACTCCGCCACATCTG
Rat TAZR	CCCTTTCTGATAGACACCATGTC
Rodent GAPDH control	Applied Biosystems Cat No.4308313

6 hr at 37°C. After DNase treatment RNA was purified with Ambion MEGAclear spin columns, then treated with Antarctic Phosphatase (New England Biolabs) for 30 min at 37°C .Treated

RNA was repurified and adjusted to 100 ng/mL working concentration with Tris-EDTA (pH 7.0).

Modified mRNA transfection were performed with RNAiMAX (Invitrogen). Transfection media was supplemented with 200 ng/ml B18R interferon inhibitor (eBioscience). After 4 hours, the transfection medium was replaced with fresh culture medium containing 200 ng/ml B18R.

Microcontact Printing

Standard soft lithography techniques were used to fabricate polydimethylsiloxane (PDMS) stamps for microcontact printing (Sylgard 184, Dow Corning Midland, MI), as previously described. Briefly, a silicon wafer was spun coat with SU-8 3005 (MicroChem Corp., Newton, MA) and selectively exposed to UV light using a photomask. After being developed, the wafer was used as a template for PDMS stamps. For the single cell studies, we used stamps with 95 μm x 13 μm rectangles. For the muscular thin film studies, we used stamps with 15 μm wide lines separated by 2 μm.

To measure sarcomere organization in single cell studies, glass coverslips (diameter 18 mm) were spun coat with PDMS and cured. PDMS stamps were coated with 50 μg/mL fibronectin (BD Biosciences, San Jose, CA) for one hour, dried, and inverted onto the coverslips after treatment in a UVO cleaner (Jelight Company Inc., Irvine, CA). Stamps were removed and the coverslips were incubated in 1% F127 Pluronic Acid (BASF, Mount Olive, NJ) for at least five minutes before rinsing with PBS and storage at 4°C. MACS-purified iPSC-CMs were plated

on microfabricated fibronectin islands for five days and transfected daily with the indicated modRNA.

Muscular Thin Film Fabrication and Experiments

Muscular thin film (MTF) chips were fabricated on 22 mm X 22 mm X 0.13-0.16 mm thick glass coverslips (Ted Pella Inc., Redding, CA). Coverslips were covered with low adhesion Scotch tape (3M, St. Paul, MN) and two rectangles of dimensions 18 mm X 5.8 mm spaced 8.6 mm apart (center to center distance) were cut into the tape with a 10.6 micron wavelength CO_2 laser prototyping system (VersaLaser 2.0, 10W, Universal Laser systems, Scottsdale, AZ). Cut rectangles were peeled using a sharp tweezer and then 10% (w/v) solution of poly(N-isopropylacrylamide), PIPAAm, (Polysciences, Inc., Warrington, PA) in 99% butanol was spun coat at 6000 rpm for 1 minute. This allowed the PIPAAm deposition within bare glass regions. The rest of the tape was then peeled off and PDMS mixed at 10:1 base to curing agent ratio was spun coat at 4000 rpm for 1 minute. PDMS-coated chips were placed in a 65°C for at least 8 hours to allow complete curing of the elastomer. Young's modulus in compression of the cured Sylgard 184 mixed in the ratio of 10:1 base to curing agent ratio was determined to be 1.52 ± 0.05 MPa (N = 18 samples, Mean ± Standard deviation) using an Instron 3342 mechanical apparatus (Instron, Norwick, MA). In the final step, two rows of cantilever outlines were cut into the elastomer within the PIPAAm rectangular regions such that the final cantilevers were 5 mm X 2 mm spaced 2.5 mm apart (center to center distance). For each batch of films, the thickness of the elastomer was measured using a profilometer (Dektak 6M, Veeco Instruments Inc., Plainview, NY) and found to be in the range of 11.4-13.4 μm.

MACS-purified iPSC-CMs were seeded on MTF constructs at a density of $10^5/cm^2$ and allowed to develop for five days, with daily transfection with modRNA as indicated. For contraction assays, MTF constructs were transferred to 37°C Tyorde's buffer solution (1.8 mM $CaCl_2$, 5 mM HEPES, 1 mM $MgCl_2$, 5.4 mM KCl, 135 mM NaCl, 0.33 mM NaH_2PO_4, and either 5 mM glucose or galactose depending on the experimental conditions, pH 7.4) and placed on the stage of a Zeiss Discovery V8 Stereo Microscope at room temperature. Tweezers were used to manually peel each thin film away from the glass coverslip as the PIPAAm layer dissolved due to the slight drop in temperature. When all films were peeled, the constructs were re-warmed to 37°C and paced with platinum field stimulation electrodes. Films were paced at 1, 2, 3, 4, and 5 Hz and their movement was recorded from above at 100 frames per second.

MTF Stress Calculation

The longitudinal planar projections of contracting MTFs were automatically detected using custom ImageJ (NIH, Bethesda, MD) software and used to calculate the radius of curvature of each film using custom MATLAB (Mathworks, Natick, MA) software, as previously described. The stress of the cell layer was determined from the radius of curvature using a modified form of Stoney's equation:

$$\sigma_{cell} = \frac{Et_s^2}{6(1-\nu^2)Rt_c(1+t_c/t_s)}, \qquad (1)$$

where σ_{cell} is the stress of the cell layer, E, ν and t_s are the Young's modulus, Poisson's ratio, and thickness of the PDMS film, respectively, R is MTF radius of curvature, and t_c is cell layer thickness. Equation (1) can be readily derived based on the theory of the cylindrical bending of thin plates and the static equilibrium of the force and torque of plate bending. Note

that the plate modulus $E/(1 - \nu^2)$ instead of the biaxial modulus $E/(1 - \nu)$, appears in equation (1) because the anisotropic contraction of the cell layer bends the PDMS film into a cylindrical shape instead of a bowl-like shape. The factor $(1 + t_c/t_s)^{-1}$ is a correction to the standard Stoney's equationwhen the thickness of the cell layer approaches that of the PDMS layer. We previously used a more comprehensive model to calculate not only the stress in the film but also the shortening of the muscle layer. For the MTFs used in this paper, the stresses calculated by these two methods are almost identical, so we chose to adopt the simpler analytical form of the modified Stoney's equation.

Stress values for the six conditions failed the Shapiro-Wilkinson test for normality and were thus statistically compared using Kruskal-Wallis One-Way ANOVA on Ranks and Dunn's method for pairwise comparisons. Tests with a p-value less than 0.05 were considered statistically significant.

Single Cell Structural Analysis

Images of single myocytes stained for sarcomeric α-actinin were analyzed using custom-designed software in ImageJ (NIH, Bethesda, MD) and MATLAB (Mathworks, Natick, MA). Images were pre-processed to highlight the filamentous structure of the cytoskeleton using a tubeness operator, which replaces each pixel in the image with the largest non-positive eigenvalue of the image Hessian matrix. To calculate the ability of single cells to spread across the microcontact printed islands, the convex hulls of sarcomeric α-actinin and fibronectin binarized immunostains were obtained and utilized to calculate cell projected surface area.

The regularity of spacing between the cytoskeletal elements that stained positive for sarcomeric α-actinin was assessed by first considering the magnitude of the oscillatory portion of the 2D Fourier transform of pre-processed and binarized immunostains:

$$F(u,v) = \Im\{f(x,y)\} = \iint_{R^2} f(x,y)e^{-j2\pi(ux+vy)}dxdy \qquad (2)$$

To fully automate the analysis and remove any user-bias (*10*), 512 radial profiles of the 2D Fourier transform were summed to obtain a 1D representation – $\Gamma(\omega_n)$ – of the 2D spectrum (blue-dots in Fig. S10 Aiii and Biii) that was further normalized so that the total area under the curve would be 1. A least square minimization was performed to find the vector of parameters γ for which the function $\tilde{\Gamma}(\omega,\gamma)$ best fit the N experimental data points:

$$\tilde{\Gamma}(\omega,\gamma) = \tilde{\Gamma}_{ap}(\omega,\gamma_{ap}) + \tilde{\Gamma}_{p}(\omega,\gamma_{p}) \qquad (3)$$

$$\gamma = \min_{\gamma}[\sum_{n=1}^{N}\left(\Gamma(\omega_n) - \tilde{\Gamma}(\omega_n,\gamma)\right)^2] \qquad (4)$$

The functional form of $\tilde{\Gamma}(\omega,\gamma)$ was composed by an aperiodic component, representing the effect of poorly developed cytoskeletal structures (black curve, Equation (5)) and a periodic component (red curves, Equation (6)) relating to periodically spaced Z-disks:

$$\tilde{\Gamma}_{ap}(\omega,\gamma_{ap}) = a + be^{-c\omega}; \gamma_{ap} = \{a,b,c\} \qquad (5)$$

$$\tilde{\Gamma}_{p}(\omega,\gamma_{p}) = \sum_{k=1}^{5} a_k e^{-\left(\frac{\omega-k\omega_0}{\delta_k}\right)^2}; \gamma_{ap} = \{\omega_0, a_k, \delta_k\} \qquad (6)$$

In agreement with Fourier analysis, the periodic component was approximated with a series of Gaussian peaks localized at integer multiples of the spatial frequency ($\omega_0 = 1/r_0$) associated with the sarcomere length ($r_0 \sim 2$ μm).

The area under the peaks of the periodic component was taken as a metric of structural organization and named sarcomere organization: that is, the organization increases as more sarcomeric α-actinin positive elements become localized in the Z-disks, at a distance $\sim r_0$. We normalized all sarcomere organization values to the maximum value observed across all single cells before plotting.

Statistics

Unless otherwise noted, results are expressed as mean ± SEM. Group means for metabolic assays were compared using Welch's t-test. For the sarcomere assembly and MTF assays, values were first tested for normality (Shapiro-Wilkinson) and only then represented as mean ± standard error of the mean. Statistical comparison between the four groups was conducted through 1-way ANOVA test followed by post-hoc Student-Newman-Keuls pairwise comparison. Differences were deemed statistically significant when statistical tests returned a p-value lower than 0.05.

References

1. F. Lan *et al.*, Abnormal calcium handling properties underlie familial hypertrophic cardiomyopathy pathology in patient-specific induced pluripotent stem cells., *Cell stem cell* **12**, 101–13 (2013).

2. C. Kim *et al.*, Studying arrhythmogenic right ventricular dysplasia with patient-specific iPSCs., *Nature* **494**, 105–10 (2013).

3. X. Carvajal-Vergara *et al.*, Patient-specific induced pluripotent stem-cell-derived models of LEOPARD syndrome., *Nature* **465**, 808–12 (2010).

4. K. Karikó, M. Buckstein, H. Ni, D. Weissman, Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA., *Immunity* **23**, 165–75 (2005).

5. A. Grosberg, P. W. Alford, M. L. McCain, K. K. Parker, Ensembles of engineered cardiac tissues for physiological and pharmacological study: heart on a chip., *Lab on a chip* **11**, 4165–73 (2011).

6. S. Bione, P. D'Adamo, E. Maestrini, A novel X-linked gene, G4. 5. is responsible for Barth syndrome, *Nature …* (1996) (available at http://www.nature.com/ng/journal/v12/n4/abs/ng0496-385.html).

7. R. H. Houtkooper *et al.*, The enigmatic role of tafazzin in cardiolipin metabolism., *Biochimica et biophysica acta* **1788**, 2003–14 (2009).

8. L. Warren *et al.*, Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA., *Cell stem cell* **7**, 618–30 (2010).

9. M. a Laflamme *et al.*, Cardiomyocytes derived from human embryonic stem cells in pro-survival factors enhance function of infarcted rat hearts., *Nature biotechnology* **25**, 1015–24 (2007).

10. M. Schlame *et al.*, Phospholipid abnormalities in children with Barth syndrome, *Journal of the American College of Cardiology* **42**, 1994–1999 (2003).

11. W. Kulik *et al.*, Bloodspot assay using HPLC-tandem mass spectrometry for detection of Barth syndrome., *Clinical chemistry* **54**, 371–8 (2008).

12. P. Mali *et al.*, RNA-guided human genome engineering via Cas9., *Science (New York, N.Y.)* **339**, 823–6 (2013).

13. J. R. Hom *et al.*, The permeability transition pore controls cardiac mitochondrial maturation and myocyte differentiation., *Developmental cell* **21**, 469–78 (2011).

14. a M. Gerdes, J. M. Capasso, Structural remodeling and mechanical dysfunction of cardiac myocytes in heart failure., *Journal of molecular and cellular cardiology* **27**, 849–56 (1995).

15. A. W. Feinberg *et al.*, Muscular thin films for building actuators and powering devices., *Science (New York, N.Y.)* **317**, 1366–70 (2007).

16. I. J. Domian *et al.*, Generation of functional ventricular heart muscle from mouse ventricular progenitor cells., *Science (New York, N.Y.)* **326**, 426–9 (2009).

17. P. W. Alford, A. W. Feinberg, S. P. Sheehy, K. K. Parker, Biohybrid thin films for measuring contractility in engineered cardiovascular muscle., *Biomaterials* **31**, 3613–21 (2010).

Chapter 6

Conclusion and future prospects

Conclusion

Here, I summarize the work described in this thesis and provide avenues for future investigation. In Chapter 2, I described re-TALENs as an improved tool for genome editing and created an optimized pipeline for generating scarlessly edited hiPSCs within 3 weeks. This pipeline includes: 1) a simplified re-coded TALE generation protocol that that facilitates genome editing tool synthesis and enables the production of functional lentivirus particles; 2) a bioinformatics package and platform, GEAS, to assess the frequency of genome editing events that is 400 times more sensitive compared with currently existing methods; and 3) an efficient genotype screening method of monoclonal hiPSCs. In Chapter 3, I described reprogramming the CRISPR system into a facile and effective tool for human cell engineering. The design and construction simplicity of this tool empowers researchers with unprecedented flexibility to conduct human genome engineering. To compare the performance of the new CRISPR system with existing genome editing tools, we compared the efficiency and specificity of CRISPRs relative to reTALENs. To enhance the performance of CRISPR, we devised a double-nickase system to improve specificity and mitigate off-target effects. To investigate safer genome editing, we created targeted deaminases as described in Chapter 4. This novel class of genome-editing tools introduces mutations in the human genome without incurring DSBs or nicks, and therefore can be applied in the practice of multiplexed-genome targeting where other nucleases-based tools may be toxic. To demonstrate the utility of our genome editing tools, in Chapter 5, we demonstrated the use of our genome editing tools in combination with the “heart-on-a-chip” technology to model cardiovascular disease and study the pathogenesis.

Resolving Specificity Issues

To fully realize the potential of genome editing tool in basic research and clinic medicine, a task of utmost importance is to improve specificity and mitigate deleterious off-target effects. The generation of off-target DSBs can generate undesirable mutations, thus confounding biological studies and impeding the use of gene-editing tools in clinical practice.

There are two aspects of specificity issues that require resolution: first, genomic DSBs introduced by a customized nuclease needs to be sequence-specific; second, the genomic changes following occurrence of a DSB should be dictated by the donor DNA -- not the non-specific NHEJ products.

To improve the DSB sequence specificity, we propose three broad strategies: 1) to determine the target bias of different type of nucleases and judiciously choose the target site; 2) to evolve genome-editing nucleases to have higher specificity; 3) to further engineer obligate cooperativity.

Targeted nucleases, including ZF, TALE, and CRISPR, can tolerant one to multiple mismatches in their binding sites. However, we and others have observed dramatic content difference so that the mismatches are tolerant to varied degrees depending on the position (*1*, *2*). Additional experiments are needed to elucidate the rules governing the position-dependent specificity, which would provide us with guidelines to computationally predict the optimal targeting sites with minimal off-target potential at gene of interest.

Second, directed evolution might be utilized to improve Cas9 specificity to a level sufficient to completely preclude off-target activity. Such a project is likely to require extensive modifications to the Cas9 protein. As such, novel methods permitting many rounds of evolution

in a short timeframe (*3*) may be warranted. For more detailed reviews of CRISPR systems, see references (*4*, *5*).

Third, efforts have been made to engineer obligated dimers to increase the specificity of customized nucleases, such as ZFN/TALEN in which the FokI nuclease domain only functions as a dimer. Our group recently reported off-set Cas9-nickases as a new strategy to mitigate the off-target issue of CRISPR system (*1*). Moving forward, more engineering is needed to decrease the affinity of nuclease monomers and enhancer cooperative effects between the monomers.

To ensure specific genomic changes at DSBs, coupling of genomic cutting and HR is critical to promote HR and disfavor undesirable non-specific NHEJ events. There may exist multiple ways in which DSB and HR can be coupled to mitigate non-specific NHEJ events. An example of spatial coupling would be conjugating the DNA donor with the gRNA of CRISPR to ensure the availability of DNA donor near the CRISPR cutting site. An example of temporal coupling may involve fusing customized nucleases with a proper cyclin domain to synchronize the expression of nucleases in G1/S of the cell cycle, during which HR is most active, thus maximizing the likelihood of an HR event relative to the likelihood of an NHEJ event (*6*). Third, DSB generation and HR can be enzymatically coupled as illustrated by the mechanism of natural recombinases (*7*). The theoretically appealing, engineered recombinases with fully programmable specificity may serve as a better alternative to nucleases for introducing specific changes in the genome.

Genome editing beyond cellular level: Patients-on-a-ChiP and *in vivo* editing

In chapter 5, we described our effort of combining engineered hiPSC with “heart on-a-ChiP” to study the pathogenesis of a certain cardiovascular disease. Currently, several organ-on-

chips are being designed in an effort to reproduce key structural, functional and mechanical properties of human organs (*8*, *9*), including lung-on-a-ChiP, heart-on-a-Chip, kidney-on-a-Chip, artery-on-a-Chip. Such technology can be extended to reconstruct liver, neuronal tissue, and gastrointestinal organs. In addition, researchers are working towards building a multi-channel 3D microfluidic system that mimics multiple organs in the whole body (*10*). We envision that the human genome editing technology, generating stem cells with designated genome type, combined with organ-on-Chips technology, will expand the capacity of cell culture models and provide low-cost alternatives to animal and clinical studies for drug screening and toxicology applications for particular diseases.

Furthermore, genome editing can be applied in higher levels of life-forms, from single cell to organic and to organismic levels. Combining efficient genetic editing technologies with specific gene delivery methods will enable the application of these tools as direct therapeutic approaches *in vivo*.

Beyond genome editing: multiple levels of cellular manipulation

Over the course of last decade, the rapid innovations of genome editing tools have dramatically enhanced our ability to manipulate the primary genomic DNA sequence. With powerful tools to recognize sequence of interest, we suggest that other levels of cellular manipulation can be also achieved, including 1) chromosomal 3D structure modulation, 2) epigenetic information recoding, and 3) transcriptional regulation.

Recent genomic studies have shown that the unique, higher-order genome structure profoundly influences transcriptional regulation (*11*). Theoretically, we can design DNA-binding proteins with multiple independent domains fused by linkers with each one recognizing distinct

region in the genome. This hypothetical chimeric protein would serve as a scaffold to bring designated region into vicinity. Such a tool can be used to investigate the still elusive functionalities of chromosomal spatial arrangement as well as to manipulate long-range gene regulation.

Second, in principle, recruitment by DNA binding protein of any of the major chromatin remodeling complexes, including SWI/SNF, histone acetylases and deacetylases, methylases and demethylases, kinases and phosphatases, DNA methylases and demethylases could potentially facilitate targeted reprogramming of chromatin modification endogenous loci. If successful, these capabilities will transform our ability to investigate the nature of epigenetic control and to engineer long-lasting gene expression changes.

Finally, targeted recruitment of transcriptional activators, such as VP64, or suppressors, such as SID and KRAB domain, by DNA-binding proteins could be deployed to directly modulate the transcription of endogenous genes to a desired level of activity. For example, we can deliver the transcriptional suppressor of HMG-CoA into liver to achieve a desired decrease in the synthesis of cholesterol, thus theoretically treating hypercholesterolemia in a transcriptional manner, akin to the use of small-molecule drugs used to modulate cholesterol synthesis. Due to the design simplicity of customized transcription regulators, we can apply the same principle to address unmet needs of medicine. For example, there is no effective drug to target mutated κ-ras, which accounts for most of the untreatable non-small-cell-lung cancers. In principle, potent κ -ras expression suppressor can be administrated into lung tissue through inhalation, thus delaying the progression of symptoms.

Synthetic life

This thesis describes the creation of a repertoire of human genome editing tools that will significantly enhance disease modeling, functional genomics and gene therapy. The convergence of both effective genome "reading" capabilities with genome "editing" technologies will be transformative for biologists dissecting the genetic determinants of disease and for clinicians who aim to deliver customized cell therapies for their patients. These advances will be complemented by strengthening genome "writing" capability to insert new genetic programs into the cell by assembling new sequences *de novo*. For example, we will be able to engineer muscle cells to secrete insulin in a highly regulated fashion, behaving as functional pancreatic beta cells. The potential to insert new genetic programs into any therapeutically desirable target will likely provide us new venues to combat diseases, enhance function and longevity.

References

1. P. Mali *et al.*, CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering, *Nature Biotechnology* (2013), doi:10.1038/nbt.2675.

2. P. D. Hsu *et al.*, DNA targeting specificity of RNA-guided Cas9 nucleases., *Nature biotechnology* , 1–8 (2013).

3. K. M. Esvelt, J. C. Carlson, D. R. Liu, A system for the continuous directed evolution of biomolecules., *Nature* **472**, 499–503 (2011).

4. G. Gasiunas, R. Barrangou, P. Horvath, V. Siksnys, Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria., *Proceedings of the National Academy of Sciences of the United States of America* **109**, E2579–86 (2012).

5. B. Wiedenheft, S. H. Sternberg, J. a Doudna, RNA-guided genetic silencing systems in bacteria and archaea., *Nature* **482**, 331–8 (2012).

6. M. Shrivastav, L. P. De Haro, J. a Nickoloff, Regulation of DNA double-strand break repair pathway choice., *Cell research* **18**, 134–47 (2008).

7. W. R. a Brown, N. C. O. Lee, Z. Xu, M. C. M. Smith, Serine recombinases as tools for genome engineering., *Methods (San Diego, Calif.)* **53**, 372–9 (2011).

8. D. Huh, G. a Hamilton, D. E. Ingber, From 3D cell culture to organs-on-chips., *Trends in cell biology* **21**, 745–54 (2011).

9. D. Huh *et al.*, Reconstituting organ-level lung functions on a chip., *Science (New York, N.Y.)* **328**, 1662–8 (2010).

10. M. B. Esch, T. L. King, M. L. Shuler, The role of body-on-a-chip devices in drug and toxicity studies., *Annual review of biomedical engineering* **13**, 55–72 (2011).

11. R. E. Thurman *et al.*, The accessible chromatin landscape of the human genome., *Nature* **489**, 75–82 (2012).

Appendix A

Optimization of scarless human stem cell genome editing

Nucleic Acids Research Advance Access published July 31, 2013

Nucleic Acids Research, 2013, 1–13
doi:10.1093/nar/gkt555

Optimization of scarless human stem cell genome editing

Luhan Yang[1,2], Marc Guell[1], Susan Byrne[1], Joyce L. Yang[1,2], Alejandro De Los Angeles[3], Prashant Mali[1], John Aach[1], Caroline Kim-Kiselak[2], Adrian W Briggs[1], Xavier Rios[1], Po-Yi Huang[1,4], George Daley[3] and George Church[1,5,*]

[1]Department of Genetics, Harvard Medical School, Boston, 02115 MA, USA, [2]Biological and Biomedical Sciences Program, Harvard Medical School, Boston, 02115 MA, USA, [3]Children's Hospital, Boston, 02115 MA, USA, [4]Chemistry and Chemical Biology program, Harvard, 02138 Cambridge, MA, USA and [5]Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, 02138 MA, USA

Received March 25, 2013; Revised May 17, 2013; Accepted May 28, 2013

ABSTRACT

Efficient strategies for precise genome editing in human-induced pluripotent cells (hiPSCs) will enable sophisticated genome engineering for research and clinical purposes. The development of programmable sequence-specific nucleases such as Transcription Activator-Like Effectors Nucleases (TALENs) and Cas9-gRNA allows genetic modifications to be made more efficiently at targeted sites of interest. However, many opportunities remain to optimize these tools and to enlarge their spheres of application. We present several improvements: First, we developed functional re-coded TALEs (reTALEs), which not only enable simple one-pot TALE synthesis but also allow TALE-based applications to be performed using lentiviral vectors. We then compared genome-editing efficiencies in hiPSCs mediated by 15 pairs of reTALENs and Cas9-gRNA targeting *CCR5* and optimized ssODN design in conjunction with both methods for introducing specific mutations. We found Cas9-gRNA achieved 7–8× higher non-homologous end joining efficiencies (3%) than reTALENs (0.4%) and moderately superior homology-directed repair efficiencies (1.0 versus 0.6%) when combined with ssODN donors in hiPSCs. Using the optimal design, we demonstrated a streamlined process to generated seamlessly genome corrected hiPSCs within 3 weeks.

INTRODUCTION

Precise genome editing in human-induced pluripotent cells (hiPSCs) will enable functional studies of human genetic variation and enhance the potential use of hiPSCs for regenerative medicine. Currently, genome editing via sequence-specific nucleases represents the most efficient way to precisely edit human cell genomes (1 3). A nuclease-mediated double-stranded DNA (dsDNA) break in the genome can be repaired by two main mechanisms (4): non-homologous end joining (NHEJ), which frequently results in the introduction of non-specific insertions and deletions (indels), or homology-directed repair (HDR), which incorporates a homologous strand as a repair template. When a sequence-specific nuclease is delivered along with a homologous donor DNA construct containing the desired mutations, gene targeting efficiencies are increased by 1000-fold compared with just the donor construct alone (5). Thus, the development of programmable nucleases has greatly facilitated the practice of targeted genome engineering.

Despite large advances in gene editing tools, many challenges and questions remain regarding the use of custom-engineered nucleases in hiPSC engineering. First, despite their design simplicity, Transcription Activator-Like Effectors Nucleases (TALENs) target particular DNA sequences with tandem copies of Repeat Variable Diresidue (RVD) domains (6). Although the modular nature of RVDs simplifies TALEN design, their repetitive sequences complicate methods for synthesizing their DNA constructs (7 10) and also impair their use with lentiviral gene delivery vehicles, most likely by causing sequence instabilities (11).

Next, we sought to improve the ease and sensitivity of current detection methods for assessing genome editing. In

*To whom correspondence should be addressed. Tel: +1 617 432 3675; Fax: +1 617 432 6513; Email: gchurch@genetics.med.harvard.edu

current practice, NHEJ and HDR are frequently evaluated using separate assays. Mismatch-sensitive endonuclease assays (12) are often used for assessing NHEJ, but the quantitative accuracy of this method is variable, and the sensitivity is limited to NHEJ frequencies greater than ~3% (12). Meanwhile, HDR is frequently assessed by cloning and sequencing, a completely different and often cumbersome procedure. Sensitivity is still an issue because, although high editing frequencies on the order of 50% are frequently reported for some cell types, such as U2OS and K562 (10,13), frequencies are generally lower in hiPSCs (14). Recently, high editing frequencies have been reported in hiPSC and hESC using TALENs (15) and even higher frequencies with the CRISPR Cas9-gRNA system (16 19). However, editing rates at different sites appear to vary widely (17), and editing is sometimes not detectable at all at some sites (20). Moreover, although the recent successes in editing hiPSC genomes with TALENs and Cas9 are striking, genome editing using these tools has not yet been systematically explored and compared. To come to a fuller understanding of these issues and optimize inefficiencies will require simple and efficient collection and analysis of NHEJ and HDR rates at large numbers of sites using tools that accurately capture low as well as high rates. To this end, we developed a robust and user-friendly package using next generation sequencing to screen HR and NHEJ events in hiPSCs together.

As a demonstration of how our improved synthesis method for TALEs, and our genome editing assessment tool, can expedite data gathering, analysis and optimization, we used these tools to compare reTALEN and Cas9 efficiencies in hiPSCs at 15 sites near the *CCR5* locus. As with TALEN and Cas9 editing of hiPSCs, generally, use of ssODNs as DNA donors has been reported (21,22), but the optimal design and scope of ssODNs for this purpose have not been systematically explored. We then used our tools to optimize the design of ssODNs used as donors for scarless genome engineering.

Another area for improvement in editing procedures for hiPSC relates to the clonal isolation of the hiPSCs themselves, an operation that is difficult in part because hiPSC are difficult to grow out from isolated single cells because in the absence of appropriate cell-to-cell contacts with other hiPSCs or feeder cells. However, procedures that improve clonal hiPSC isolation have recently been reported (23), and we adapted these to integrate with the other procedures we report here. Taken all together, we demonstrate that it is possible to obtain clonal, precisely genome-edited hiPSCs within 3 weeks, including within this the amount of time required to synthesize optimal reagents and perform rapid prospective screening of target events.

MATERIALS AND METHODS

gRNA assembly

We incorporated 19 bp of the selected target sequence (i.e. 5′-N_{19} of 5′-N_{19}-NGG-3′) into two complementary 100 mer oligonucleotides (TTCTTGGCTTTATATATCTTG TGGAAAGGACGAAACACCGN19GTTTTAGAGCT AGAAATAGCAAGTTAAAATAAGGCTAGTCC). Each 100 mer oligonucleotide was suspended at 100 mM in water, mixed with equal volume and annealed in thermocycle machine (95°C, 5 min; Ramp to 4°C, 0.1°C/s). To prepare the destination vector, we linearized the gRNA cloning vector (Addgene plasmid ID 41824, Supplementary Sequence S3) using AfIII and purified the vector through purification. We carried out the (10 µl) gRNA assembly reaction with 10 ng annealed 100 bp fragment, 100 ng destination backbone, 1× Gibson assembly reaction mix (New England Biolabs) at 50°C for 30 min, and reaction can be processed directly for bacterial transformation to colonize individual assemblies.

re-TALEs design and assembly

re-TALEs were optimized at different levels to facilitate assembly and improve expression. re-TALE DNA sequences were first co-optimized for a human codon-usage and low mRNA folding energy at the 5′ end (GeneGA, Bioconductor). The obtained sequence was evolved through several cycles to eliminate repeats (direct or inverted) longer than 11 bp (Supplementary Figure S8). In each cycle, synonymous sequences for each repeat are evaluated. Those with the largest hamming distance to the evolving DNA are selected. The sequence of one of re-TALE possessing 16.5 monomers is listed in Supplementary Sequence S1.

re-TALE dimer blocks encoding two RVDs (Supplementary Figure S2A) were generated by two rounds of PCR under standard Kapa HIFI (KPAP) PCR conditions, in which the first round of PCR introduced the RVD coding sequence and the second round of PCR generated the entire dimer blocks with 36 bp overlaps with the adjacent blocks. PCR products were purified using QIAquick 96 PCR Purification Kit (QIAGEN), and the concentrations were measured by Nano-drop. The primer and template sequences are listed in Supplementary Tables S1 and S2.

re-TALENs and re-TALE-TF destination vectors were constructed by modifying the TALE-TF and TALEN cloning backbones (24). We re-coded the 0.5 RVD regions on the vectors and also incorporated SapI cutting site at the designated re-TALE cloning site. The sequences of re-TALENs and re-TALE-TF backbones are listed in Supplementary Sequence S2. Plasmids can be pretreated with SapI (New England Biolabs) with manufacturer recommended conditions and purified with QIAquick PCR purification kit (QIAGEN).

We carried out the (10 µl) one-pot TALE Single-incubation Assembly (TASA) assembly reaction with 200 ng of each block, 500 ng of destination backbone, 1× TASA enzyme mixture [2U SapI, 100 U Ampligase (Epicentre), 10 mU T5 exonuclease (Epicentre), 2.5U Phusion DNA polymerase (New England Biolabs)] and 1× isothermal assembly reaction buffer as described before (25) [5% PEG-8000, 100 mM Tris HCl (pH 7.5), 10 mM MgCl2, 10 mM DTT, 0.2 mM each of the four dNTPs and 1 mM NAD]. Incubations were performed at 37°C for 5 min and 50°C for 30 min. TASA assembly

reaction can be processed directly for bacterial transformation to colonize individual assemblies. The efficiency of obtaining full-length construct is ~20% with this approach. Alternatively, >90% efficiency can be achieved by three-steps assembly. First, 10 µl of re-TALE assembly reactions were performed with 200 ng of each block, 1× re-TALE enzyme mixture (100 U Ampligase, 12.5 mU T5 exonuclease, 2.5 U Phusion DNA polymerase) and 1× isothermal assembly buffer at 50°C for 30 min, followed by standardized Kapa HIFI PCR reaction, agarose gel electrophoresis and QIAquick Gel extraction (Qiagen) to enrich the full-length re-TALEs. In all, 200 ng of re-TALE amplicons can then be mixed with 500 ng of Sap1-pre-treated destination backbone, 1× re-TALE assembly mixture and 1× isothermal assembly reaction buffer and incubated at 50°C for 30 min. The re-TALE final assembly reaction can be processed directly for bacterial transformation to colonize individual assemblies. Additional notes of the assembly methods can be found in Supplementary Note S1.

Cell line and cell culture

PGP1 iPS cells were maintained on Matrigel (BD Biosciences)-coated plates in mTeSR1 (Stemcell Technologies). Cultures were passaged every 5 7 days with TrypLE Express (Invitrogen). The 293 T and 293FT cells were grown and maintained in Dulbecco's modified Eagle's medium (DMEM, Invitrogen) high glucose supplemented with 10% fetal bovine serum (Invitrogen), penicillin/streptomycin (pen/strep, Invitrogen) and non-essential amino acids (Invitrogen). K562 cells were grown and maintained in RPMI (Invitrogen) supplemented with 10% fetal bovine serum (Invitrogen 15%) and penicillin/streptomycin (pen/strep, Invitrogen). All cells were maintained at 37°C and 5% CO2 in a humidified incubator.

We established a stable 293T cell line for detecting HDR efficiency as described before (26). Specifically, the reporter cell lines bear genomically integrated GFP-coding sequences disrupted by the insertion of a stop codon and a 68 bp genomic fragment derived from the AAVS1 locus.

Test of reTALENs activity

We seeded 293 T reporter cells at densities of 2 × 105 cells per well in 24-well plate and transfected them with 1 µg of each re-TALENs plasmid and 2 µg DNA donor plasmid using Lipofectamine 2000 following the manufacturer's protocols. Cells were harvested using TrypLE Express (Invitrogen) ~18 h after transfection and resuspended in 200 µl of media for flow cytometry analysis using an LSRFortessa cell analyzer (BD Biosciences). The flow cytometry data were analyzed using FlowJo (FlowJo). At least 25 000 events were analyzed for each transfection sample. For endogenous AAVS1 locus targeting experiment in 293 T, the transfection procedures were identical as described earlier in the text, and we conducted puromycin selection with drug concentration at 3 µg/ml 1 week after transfection.

Functional lentivirus generation assessment

The lentiviral vectors were created by standard PCR and cloning techniques. The lentiviral plasmids were transfected by Lipofectamine 2000 with Lentiviral Packaging Mix (Invitrogen) into cultured 293FT cells (Invitrogen) to produce lentivirus. Supernatant was collected 48 and 72 h post-transfection, sterile filtered and 100 µl of filtered supernatant was added to 5×10^5 fresh 293 T cells with polybrene. Lentivirus titration was calculated based on the following formula: virus titration = (percentage of GFP+ 293 T cell × initial cell numbers under transduction)/(the volume of original virus collecting supernatant used in the transduction experiment). To test the functionality of lentivirus, 3 days after transduction, we transfected lentivirus transduced 293 T cells with 30 ng of plasmids carrying mCherry reporter and 500 ng of pUC19 plasmids using Lipofectamine 2000 (Invitrogen). Cell images were analyzed using Axio Observer Z.1 (Zeiss) 18 h after transfection and harvested using TrypLE Express (Invitrogen) and resuspended in 200 µl of media for flow cytometry analysis using LSRFortessa cell analyzer (BD Biosciences). The flow cytometry data were analyzed using BD FACSDiva (BD Biosciences).

Test of re-TALENs and Cas9-gRNA genome editing efficiency

PGP1 iPSCs were cultured in Rho kinase (ROCK) inhibitor Y-27632 (Calbiochem) 2 h before nucleofection. Transfections were done using P3 Primary Cell 4D-Nucleofector X Kit (Lonza). Specifically, cells were harvested using TrypLE Express (Invitrogen), and 2×10^6 cells were resuspended in 20 µl of nucleofection mixture containing 16.4 µl of P3 Nucleofector solution, 3.6 µl of supplement, 1 µg of each re-TALENs plasmid or 1 µg of Cas9 and 1 µg of gRNA construct, 2 µl of 100 µM ssODN. Subsequently, we transferred the mixtures to 20 µl of Nucleocuvette strips and conducted nucleofection using CB150 program. Cells were plated on Matrigel-coated plates in mTeSR1 medium supplemented with ROCK inhibitor for the first 24 h. For endogenous AAVS1 locus-targeting experiment with dsDNA donor, we used the identical procedure except we used 2 µg of dsDNA donor, and we supplement the mTeSR1 media with puromycin at the concentration of 0.5 µg/ml 1 week after transfection.

The information of reTALENs, gRNA and ssODNs used in this study are listed in Supplementary Tables S3 and S6.

Amplicon library preparation of the targeting regions

Cells were harvested 6 days after nucleofection and 0.1 µl of prepGEM tissue protease enzyme (ZyGEM) and 1 µl of prepGEM gold buffer (ZyGEM) were added to 8.9 µl of the 2 5×10^5 cells in the medium. In all, 1 µl of the reactions were then added to 9 µl of PCR mix containing 5 µl 2 × KAPA Hifi Hotstart Readymix (KAPA Biosystems) and 100 nM corresponding amplification primer pairs. Reactions were incubated at 95°C for 5 min followed by 15 cycles of 98°C, 20 s; 65°C, 20 s and 72°C, 20 s. To add

the Illumina sequence adaptor, 5 µl of reaction products were then added to 20 µl of PCR mix containing 12.5 µl of 2 × KAPA HIFI Hotstart Readymix (KAPA Biosystems) and 200 nM primers carrying Illumina sequence adaptors. Reactions were incubated at 95°C for 5 min followed by 25 cycles of 98°C, 20 s; 65°C, 20 s and 72°C, 20 s. PCR products were purified by QIAquick PCR purification kit, mixed at roughly the same concentration and sequenced with MiSeq Personal Sequencer. All the PCR primers can be found in the Supplementary Table S5.

Genome editing assessment system

We wrote a pipeline to analyze the genome engineering data. This pipeline is integrated in one single Unix module, which uses different tools such as R, BLAT and FASTX Toolkit.

Barcode splitting: Groups of samples were pooled together and sequenced using MiSeq 150 bp paired end (PE150) (Illumina Next Gen Sequencing) and later separated based on DNA barcodes using FASTX Toolkit.

Quality filtering: We trimmed nucleotides with lower sequence quality (phred score <20). After trimming, reads shorter than 80 nt were discarded.

Mapping: We used BLAT to map the paired reads independently to the reference genome and we generated .psl files as output.

Indel calling: We defined indels as the full-length reads containing two blocks of matches in the alignment. Only reads following this pattern in both paired end reads were considered. As a quality control, we required the indel reads to possess minimal 70 nt matching with the reference genome and both blocks to be at least 20 nt long. Size and position of indels were calculated by the positions of each block to the reference genome. Non-homologous end joining (NHEJ) has been estimated as the percentage of reads containing indels [see Equation (1)]. The majority of NHEJ event have been detected at the targeting site vicinity.

Homology-directed recombination (HDR) efficiency: Pattern matching (grep) within a 12 bp window centering over DSB was used to count specific signatures corresponding to reads containing the reference sequence, modifications of the reference sequence (2 bp intended mismatches) and reads containing only 1 bp mutation within the 2 bp intended mismatches [see Equation (1)].

Equation 1. Estimation of NHEJ and HDR

A = reads identical to the reference: XXXXXABX XXXX

B = reads containing 2 bp mismatch programed by ssODN: XXXXXabXXXXX

C = reads containing only 1 bp mutation in the target site: such as XXXXXaBXXXXX or XXXXXAbXXXXX

D = reads containing indels as described above

$$\text{NHEJ efficiency} = \left(100 \times \frac{D}{A+B+C+D}\right)\%$$

$$\text{HDR efficiency} = \left(100 \times \frac{B}{A+B+C+D}\right)\%$$

The statistic analysis of the GEAS can be found in Supplementary Note S2.

Genotype screening of colonized hiPSCs

Human iPS cells on feeder-free cultures were pre-treated with mTesr-1 media supplemented with SMC4 (5 uM thiazovivin, 1 uM CHIR99021, 0.4 uM PD0325901, 2 uM SB431542) (23) for at least 2 h before fluorescence-activated cell sorting (FACS) sorting. Cultures were dissociated using Accutase (Millipore) and resuspended in mTesr-1 media supplemented with SMC4 and the viability dye ToPro-3 (Invitrogen) at concentration of 1 2 × 107 /ml. Live hiPS cells were single-cell sorted using a BD FACSAria II SORP UV (BD Biosciences) with 100 µm nozzle under sterile conditions into 96-well plates coated with irradiated CF-1 mouse embryonic fibroblasts (Global Stem). Each well contained hES cell medium (27) with 100 ng/ml recombinant human basic Fibroblast Growth Factor (Millipore) supplemented with SMC4 and 5 µg/ml fibronectin (Sigma). After sorting, plates were centrifuged at 70*g* for 3 min. Colony formation was seen 4 days post sorting, and the culture media was replaced with hES cell medium with SMC4. SMC4 can be removed from hES cell medium 8 days after sorting.

A few thousand cells were harvested 8 days after FACS and 0.1 µl of prepGEM tissue protease enzyme (ZyGEM) and 1 µl of prepGEM gold buffer (ZyGEM) were added to 8.9 µl of cells in the medium. The reactions were then added to 40 µl of PCR mix containing 35.5 ml of platinum 1.1× Supermix (Invitrogen), 250 nM of each dNTP and 400 nM primers. Reactions were incubated at 95°C for 3 min followed by 30 cycles of 95°C, 20 s; 65°C, 30 s and 72°C, 20 s. Products were Sanger sequenced using either one of the PCR primers (Supplementary Table S5), and sequences were analyzed using DNASTAR (DNASTAR).

Immunostaining and teratoma assays of hiPSCs

Cells were incubated in the KnockOut DMEM/F-12 medium at 37°C for 60 min using the following antibody: Anti-SSEA-4 PE (Millipore) (1: 500 diluted); Tra-1-60 (BD Pharmingen) (1:100 diluted). After the incubation, cells were washed three times with KnockOut DMEM/F-12 and imaged on the Axio Observer Z.1 (ZIESS).

To conduct teratoma formation analysis, we harvested human iPSCs using collagenase type IV (Invitrogen) and resuspended the cells into 200 µl of Matrigel and injected intramuscularly into the hind limbs of Rag2gamma knockout mice. Teratomas were isolated and fixed in formalin between 4 and 8 weeks after the injection. The teratomas were subsequently analyzed by hematoxylin and eosin staining.

RESULTS

ReTALENs target genomic loci effectively in human somatic and stem cells

TALEs have proven to be a powerful and easy-to-design tool for targeted genome manipulation in multiple cell

the Illumina sequence adaptor, 5 µl of reaction products were then added to 20 µl of PCR mix containing 12.5 µl of 2 × KAPA HIFI Hotstart Readymix (KAPA Biosystems) and 200 nM primers carrying Illumina sequence adaptors. Reactions were incubated at 95°C for 5 min followed by 25 cycles of 98°C, 20 s; 65°C, 20 s and 72°C, 20 s. PCR products were purified by QIAquick PCR purification kit, mixed at roughly the same concentration and sequenced with MiSeq Personal Sequencer. All the PCR primers can be found in the Supplementary Table S5.

Genome editing assessment system

We wrote a pipeline to analyze the genome engineering data. This pipeline is integrated in one single Unix module, which uses different tools such as R, BLAT and FASTX Toolkit.

Barcode splitting: Groups of samples were pooled together and sequenced using MiSeq 150 bp paired end (PE150) (Illumina Next Gen Sequencing) and later separated based on DNA barcodes using FASTX Toolkit.

Quality filtering: We trimmed nucleotides with lower sequence quality (phred score <20). After trimming, reads shorter than 80 nt were discarded.

Mapping: We used BLAT to map the paired reads independently to the reference genome and we generated .psl files as output.

Indel calling: We defined indels as the full-length reads containing two blocks of matches in the alignment. Only reads following this pattern in both paired end reads were considered. As a quality control, we required the indel reads to possess minimal 70 nt matching with the reference genome and both blocks to be at least 20 nt long. Size and position of indels were calculated by the positions of each block to the reference genome. Non-homologous end joining (NHEJ) has been estimated as the percentage of reads containing indels [see Equation (1)]. The majority of NHEJ event have been detected at the targeting site vicinity.

Homology-directed recombination (HDR) efficiency: Pattern matching (grep) within a 12 bp window centering over DSB was used to count specific signatures corresponding to reads containing the reference sequence, modifications of the reference sequence (2 bp intended mismatches) and reads containing only 1 bp mutation within the 2 bp intended mismatches [see Equation (1)].

Equation 1. Estimation of NHEJ and HDR

A = reads identical to the reference: XXXXXABX XXXX

B = reads containing 2 bp mismatch programed by ssODN: XXXXXabXXXXX

C = reads containing only 1 bp mutation in the target site: such as XXXXXaBXXXXX or XXXXXAbXXXXX

D = reads containing indels as described above

$$\text{NHEJ efficiency} = \left(100 \times \frac{D}{A+B+C+D}\right)\%$$

$$\text{HDR efficiency} = \left(100 \times \frac{B}{A+B+C+D}\right)\%$$

The statistic analysis of the GEAS can be found in Supplementary Note S2.

Genotype screening of colonized hiPSCs

Human iPS cells on feeder-free cultures were pre-treated with mTesr-1 media supplemented with SMC4 (5 uM thiazovivin, 1 uM CHIR99021, 0.4 uM PD0325901, 2 uM SB431542) (23) for at least 2 h before fluorescence-activated cell sorting (FACS) sorting. Cultures were dissociated using Accutase (Millipore) and resuspended in mTesr-1 media supplemented with SMC4 and the viability dye ToPro-3 (Invitrogen) at concentration of 1 2 × 107 /ml. Live hiPS cells were single-cell sorted using a BD FACSAria II SORP UV (BD Biosciences) with 100 µm nozzle under sterile conditions into 96-well plates coated with irradiated CF-1 mouse embryonic fibroblasts (Global Stem). Each well contained hES cell medium (27) with 100 ng/ml recombinant human basic Fibroblast Growth Factor (Millipore) supplemented with SMC4 and 5 µg/ml fibronectin (Sigma). After sorting, plates were centrifuged at 70*g* for 3 min. Colony formation was seen 4 days post sorting, and the culture media was replaced with hES cell medium with SMC4. SMC4 can be removed from hES cell medium 8 days after sorting.

A few thousand cells were harvested 8 days after FACS and 0.1 µl of prepGEM tissue protease enzyme (ZyGEM) and 1 µl of prepGEM gold buffer (ZyGEM) were added to 8.9 µl of cells in the medium. The reactions were then added to 40 µl of PCR mix containing 35.5 ml of platinum 1.1× Supermix (Invitrogen), 250 nM of each dNTP and 400 nM primers. Reactions were incubated at 95°C for 3 min followed by 30 cycles of 95°C, 20 s; 65°C, 30 s and 72°C, 20 s. Products were Sanger sequenced using either one of the PCR primers (Supplementary Table S5), and sequences were analyzed using DNASTAR (DNASTAR).

Immunostaining and teratoma assays of hiPSCs

Cells were incubated in the KnockOut DMEM/F-12 medium at 37°C for 60 min using the following antibody: Anti-SSEA-4 PE (Millipore) (1: 500 diluted); Tra-1-60 (BD Pharmingen) (1:100 diluted). After the incubation, cells were washed three times with KnockOut DMEM/F-12 and imaged on the Axio Observer Z.1 (ZIESS).

To conduct teratoma formation analysis, we harvested human iPSCs using collagenase type IV (Invitrogen) and resuspended the cells into 200 µl of Matrigel and injected intramuscularly into the hind limbs of Rag2gamma knockout mice. Teratomas were isolated and fixed in formalin between 4 and 8 weeks after the injection. The teratomas were subsequently analyzed by hematoxylin and eosin staining.

RESULTS

ReTALENs target genomic loci effectively in human somatic and stem cells

TALEs have proven to be a powerful and easy-to-design tool for targeted genome manipulation in multiple cell

lines and organisms (2,13 15, 28 30). Several strategies have been developed to assemble the repetitive TALE RVD array sequences (7 10). However, once assembled, the TALE sequence repeats remain unstable, which limits the wide utility of this tool, especially for viral gene delivery vehicles (11,31). We thus thought that complete elimination of repeats would not only enable faster and simple synthesis of extended TALE RVD arrays but also address this important post-synthesis problem.

To eliminate repeats, we computationally evolved the nucleotides sequence of TALE RVD arrays to minimize the number of sequence repeats while maintaining the amino acid composition. Re-coded TALE (Re-TALEs) encoding 16 tandem RVD DNA recognition monomers, plus the final half RVD repeat, are devoid of any 12 bp repeats (Supplementary Figure S1a). Notably, this level of recoding is sufficient to allow PCR amplification of any specific monomer or sub-section from a full-length re-TALE construct (Supplementary Figure S1b). The improved design of re-TALEs makes it possible to order them directly from gene synthesis companies using standard DNA synthesis technology (32), without incurring the additional costs or procedures associated with repeat-heavy sequences. Furthermore, the recoded sequence design also enabled us to efficiently assemble re-TALE constructs using a modified isothermal assembly reaction ('Materials and Methods' section, Supplementary Note S1, Supplementary Figure S2).

We next sought to test the function of reTALEN in comparison with the corresponding non-recoded TALEN in human cells. To this end, we used a HEK 293 cell line containing a GFP reporter cassette carrying a frame-shifting insertion as previously described (33) (Figure 1a). Delivery of TALENs or reTALENs targeting the insertion sequence, together with a promoter-less GFP donor construct, leads to DSB-induced HDR repair of the GFP cassette so that GFP repair efficiency can be used to evaluate the nuclease cutting efficiency (34). We found that reTALENs induced GFP repair in 1.4% of the transfected cells, similar to that achieved by TALENs (1.2%) (Figure 1b). We further tested the activity of reTALENs at the AAVS1 locus in PGP1 hiPSCs (Figure 1c) and successfully recovered cell clones containing specific insertions (Figure 1d and e), confirming that reTALENs are active in both somatic and pluripotent human cells.

We then confirmed that the elimination of repeats would enable us to generate functional lentivirus with a re-TALE cargo. Specifically, we packaged lentiviral particles encoding re-TALE-2 A-GFP and obtained lentiviral particles with tittering of 1.3×10^6 We then tested the activity of the re-TALE-TF encoded by viral particles by transfecting a mCherry reporter into a pool of lenti-reTALE-2 A-GFP-infected 293 T cells. The 293 T cells transduced by lenti-re-TALE-TF showed 36× reporter expression activation compared with the reporter only negative (Supplementary Figure S3a c). We further checked the sequence integrity of the re-TALE-TF in the lentiviral infected cells and detected full-length reTALEs in all 10 of the clones tested (Supplementary Figure S3d).

Comparison of ReTALEs and Cas9-gRNA efficiency in hiPSCs with GEAS

To compare the editing efficiencies of re-TALENs versus Cas9-gRNA in hiPSCs, we developed a next-generation sequencing platform to precisely pinpoint and quantify both NHEJ and HDR gene-editing events, which we refer to as Genome Editing Assessment System (GEAS). First, we designed and constructed a re-TALEN pair and a Cas9-gRNA, both targeting the upstream region of CCR5 (re-TALEN, Cas9-gRNA pair #3 in Supplementary Table S3), along with a 90 nt ssODN donor identical to the target site except for a 2 bp mismatch (Figure 2a). We then transfected the nuclease constructs and donor ssODN into hiPSCs. To precisely quantitate the gene-editing efficiency, we conducted paired-end deep sequencing on the target genomic region 3 days after transfection. HDR efficiency was measured by the percentage of reads containing the precise 2 bp mismatch. NHEJ efficiency was measured by the percentage of reads carrying indels.

Delivery of the ssODN alone into hiPSCs resulted in minimal HDR and NHEJ rates, whereas delivery of the re-TALENs and the ssODN led to efficiencies of 1.7% HDR and 1.2% NHEJ (Figure 2b). The introduction of the Cas9-gRNA with the ssODN led to 1.2% HDR and 3.4% NHEJ efficiencies. Notably, the rate of genomic deletions and insertions peaked in the middle of the spacer region between the two reTALENs binding site, but peaked 3 4 bp upstream of the protospacer associated motif (PAM) sequence of Cas9-gRNA-targeting site (Figure 2b) as would be expected from the fact that DSBs take place in these regions. We observed a median genomic deletion size of 6 bp and insertion size of 3 bp generated by the re-TALENs and a median deletion size of 7 bp and insertion of 1 bp by the Cas9-gRNA (Figure 2b), consistent with DNA lesion patterns usually generated by NHEJ (4). Several analyses of our next-generation sequencing platform revealed that GEAS can detect HDR detection rates as low as 0.007%, which is both highly reproducible (coefficient of variation between replicates = ± 15% × measured efficiency) and 400× more sensitive than most commonly used mismatch sensitive endonuclease assays (Supplementary Figure S4).

After confirming the reliability of GEAS, we next sought to test the scalability of our tools by building and assessing re-TALEN pairs and Cas9-gRNAs targeted to 15 sites at the CCR5 genomic locus (Figure 2c, Supplementary Table S3). Anticipating that editing efficiency might depend on chromatin state, these sites were selected to represent a wide range of DNaseI sensitivities (35). The nuclease constructs were transfected with the corresponding ssODNs donors (Supplementary Table S3) into PGP1 hiPSCs. Six days after transfection, we profiled the genome-editing efficiencies at these sites (Supplementary Table S4). For 13 of 15 re-TALEN pairs with ssODN donors, we detected NHEJ and HDR at levels above our statistical detection thresholds, with an average NHEJ efficiency of 0.4% and an average HDR efficiency of 0.6% (Figure 2c). In addition, a statistically significant positive correlation ($r^2 = 0.81$) was found

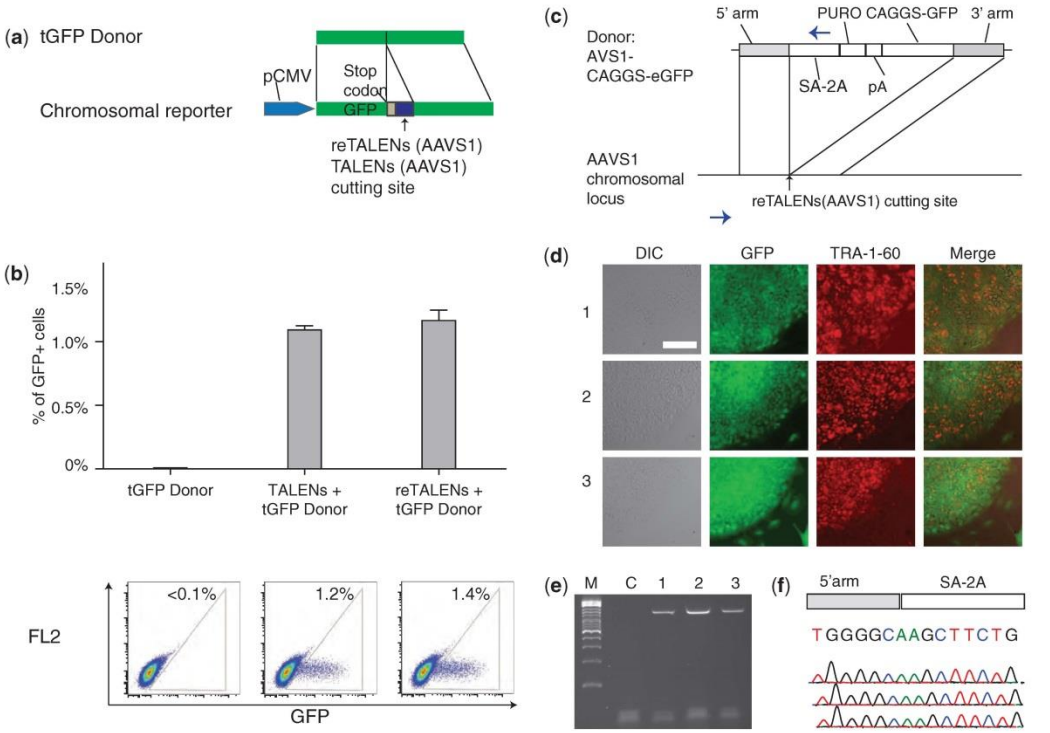


Figure 1. Functional tests of re-TALENs in human somatic and stem cells. (**a**) Schematic representation of experimental design for testing genome targeting efficiency. A genomically integrated GFP-coding sequence is disrupted by the insertion of a stop codon and a 68 bp genomic fragment derived from the AAVS1 locus (bottom). Restoration of the GFP sequence by nuclease-mediated homologous recombination with tGFP donor (top) results in GFP+ cells that can be quantitated by FACS. Re-TALENs and TALENs target identical sequences within AAVS1 fragments. (**b**) Bar graph depicting GFP+ cell percentage introduced by tGFP donor alone, TALENs with tGFP donor and re-TALENs with tGFP donor at the target locus, as measured by FACS ($N = 3$, error bar = SD). Representative FACS plots are shown later in the text. (**c**) Schematic overview depicting the targeting strategy for the native AAVS1 locus. The donor plasmid, containing splicing acceptor (SA)- 2 A (self-cleaving peptides), puromycin resistant gene (PURO) and GFP were described before (14). The locations of PCR primers used to detect successful editing events are depicted as blue arrows. (**d**) Successfully targeted clones of PGP1 hiPSCs were selected with puromycin (0.5 μg/ml) for 2 weeks. Microscopy images of three representative GFP+ clones are shown. Cells were also stained for the pluripotency markers TRA-1-60. Scale bar: 200 μm. (**e**) PCR assays performed on these the monoclonal GFP+ hiPSC clones demonstrated successful insertions of the donor cassettes at the AAVS1 site (lanes 1–3), whereas plain hiPSCs show no evidence of successful insertion (lane C). (**f**) Sanger sequencing of the PCR amplicon from the three targeted hiPSC colonies confirmed that the expected DNA bases at the genome-insertion boundary is present.

between HR and NHEJ efficiency at the same targeting loci ($P < 1 \times 10^{-4}$) (Supplementary Figure S5a), suggesting that DSB generation, the common upstream step of both HDR and NHEJ, is a rate-limiting step for reTALEN-mediated genome editing.

In contrast, all 15 Cas9-gRNA pairs showed significant levels of NHEJ and HR, with an average NHEJ efficiency of 3% and an average HDR efficiency of 1.0% (Figure 2c). In addition, a positive correlation was also detected between the NHEJ and HDR efficiency introduced by Cas9-gRNA (Supplementary Figure S5b) ($r^2 = 0.52$, $P = 0.003$), consistent with what we had observed with our reTALENs. The NHEJ efficiency achieved by Cas9-gRNA was significantly higher than that achieved by reTALENs (*t*-test, paired-end, $P = 0.02$). Interestingly, we observed a moderate but statistically significant correlation between NHEJ efficiency and the melting temperature of the gRNA targeting sequence (Supplementary Figure S5c) ($r^2 = 0.28$, $P = 0.04$), suggesting that the strength of base pairing between the gRNA and its genomic target could explain as much as 28% of the variation in the efficiency of Cas9-gRNA-mediated DSB generation. Even though Cas9-gRNA produced NHEJ levels at an average of seven times higher than the corresponding reTALEN, Cas9-gRNA only achieved HDR levels (average = 1.0%) similar to that of the corresponding reTALENs (average = 0.6%), suggesting either that the ssODN concentration at the DSB is the limiting factor for HDR or that the genomic break structure created by the Cas9-gRNA is not

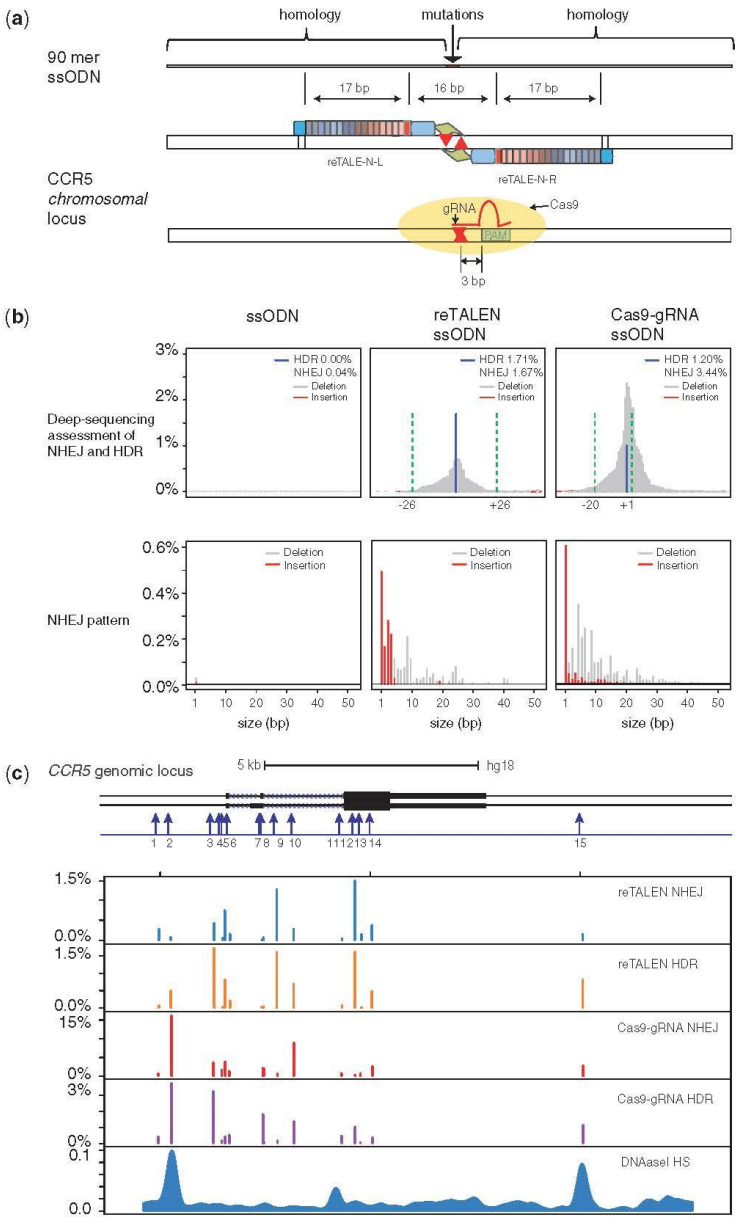


Figure 2. Comparison of reTALENs and Cas9-gRNAs genome targeting efficiency on *CCR5* in iPSCs. (**a**) Schematic representation of genome engineering experimental design. At the re-TALEN pair or Cas9-gRNA targeting site, a 90 mer ssODN carrying a 2 bp mismatch against the genomic DNA was delivered along with the reTALEN or Cas9-gRNA constructs into PGP1 hiPSCs. The cutting sites of the nucleases are depicted as red arrows in the figure. (**b**) Deep-sequencing analysis of HDR and NHEJ efficiencies for re-TALEN pairs (CCR5 #3) and ssODN, or the Cas9-gRNA and ssODN. Alterations in the genome of hiPSCs were analyzed from high-throughput sequence data by GEAS. Top: HDR was quantified from the fraction of reads that contained a 2 bp point mutation built into the center of the ssODN (blue), and NHEJ activity was quantified from the fraction of deletions (gray)/Insertions (red) at each specific position in the genome. For the reTALEN and ssODN graphs, we plot green dashed lines to mark

(continued)

favorable for effective HDR (see 'Discussion' section). Of note, within our data, we did not observe any correlation between DNaseI HS and the genome targeting efficiencies achieved by either method (Supplementary Figure S6).

Optimization of ssODN donor design for HDR

Although ssODNs have been found to be effective as donor DNA in genome editing [see earlier in the text, (21,22)], many questions remain regarding how to optimize their design. Having compared the efficiencies of reTALEN and Cas9-gRNA nucleases, we next developed strategies for the design of highly performing ssODNs in hiPSCs.

We first designed a set of ssODNs donors of different lengths (50 170 nt), all carrying the same 2 bp mismatch in the middle of the spacer region of the CCR5 re-TALEN pair #3 target sites. HDR efficiency was observed to vary with ssODN length, and an optimal HDR efficiency of ~1.8% was observed with a 90 nt ssODN, whereas longer ssODNs decreased HDR efficiency (Figure 3a). As longer homology regions improve HDR rates when dsDNA donors are used with nucleases (36), possible reasons for this result may be that ssODNs are used in an alternative genome repair process; longer ssODNs are less available to the genome repair apparatus or that longer ssODNs incur negative effects that offset any improvements gained by longer homology, compared with dsDNA donors (37). Yet, if either of the first two reasons were the case, then NHEJ rates should either be unaffected or would increase with longer ssODNs because NHEJ repair does not involve the ssODN donor. However, NHEJ rates were observed to decline along with HDR (Figure 3a), suggesting that the longer ssODNs present offsetting effects. Possible hypotheses would be that longer ssODNs are toxic to the cell (38) or that transfection of longer ssODNs saturates the DNA processing machinery, thereby causing decreased molar DNA uptake and reducing the capacity of the cells to take up or express re-TALEN plasmids.

Next, we examined how rate of incorporation of a mismatch carried by the ssODN donor varies with its distance to the DSB. To this end, we designed a series of 90 nt ssODNs all possessing the same 2 bp mismatch (A) in the center of the spacer region of re-TALEN pair #3. Each ssODN also contained a second 2 bp mismatch (B) at varying distances from the center (Figure 3b). An ssODN possessing only the center 2 bp mismatch was used as a control. Each of these ssODNs was introduced individually with re-TALEN pair #3, and the outcomes were analyzed with GEAS. We found that overall HDR as measured by the rate at which the A mismatch was incorporated (A only or A + B) decreased as the B mismatches became farther from the center (Figure 3b, Supplementary Figure S7a). The higher overall HDR rate observed when B is only 10 bp away from A may reflect a lesser need for annealing of the ssODN against genomic DNA immediately proximal to the dsDNA break.

For each distance of B from A, a fraction of HDR events only incorporated the A mismatch, whereas another fraction incorporated both A and B mismatches [Figure 3b (A only and A + B)], These two outcomes may be due to gene conversion tracts (39) along the length of the ssDNA oligo, whereby incoporation of A + B mismatches resulted from long conversion tracts that extended beyond the B mismatch, and incorporation of the A-only mismatch resulted from shorter tracts that did not reach B. Under this interpretation, we estimated a distribution of gene conversion lengths in both directions along the ssODN (Supplementary Figure S7b). The estimated distribution implies that gene conversion tracts progressively become less frequent as their lengths increase, a result similar to gene conversion tract distributions seen with dsDNA donors (39), but on a highly compressed distance scale of tens of bases for the ssDNA donor versus hundreds of bases for dsDNA donors. Consistent with this result, an experiment with a ssODN containing three pairs of 2 bp mismatches spaced at intervals of 10 nt on either side of the central 2 bp mismatch 'A's gave rise to a pattern in which A alone was incorporated 86% of the time, with multiple B mismatches incorporated at other times (Supplementary Figure S7c). Although the numbers of B only incorporation events were too low to estimate a distribution of tract lengths <10 bp, it is clear that the short tract region within 10 bp of the nuclease site predominates (Supplementary Figure S7b). Finally, in all of our experiments with single B mismatches, we see a small fraction of B-only incorporation events (0.04 0.12%) that is roughly constant across all B distances from A. The nature of these events is unclear.

Furthermore, we tested how far the ssODN donor can be placed from the re-TALEN-induced dsDNA break and

Figure 2. Continued
the outer boundary of the re-TALEN pair's binding sites, which are at positions −26 bp and +26 bp relative to the center of the two re-TALEN-binding sites. For Cas9-gRNA and ssODN graphs, the green dashed lines mark the outer boundary of the gRNA targeting site, which are at positions −20 and −1 bp relative to the Protospacer Associated Motif sequence. Bottom: Deletion/Insertion size distribution in hiPSCs analyzed from the entire NHEJ population with treatments indicated earlier in the text. (c) The genome-editing efficiency of re-TALENs and Cas9-gRNAs targeting *CCR5* in PGP1 hiPSCs. Top: schematic representation of the targeted genome-editing sites in *CCR5*. The 15 targeting sites are illustrated by blue arrows later in the text. For each site, cells were co-transfected with a pair of re-TALENs and their corresponding ssODN donor carrying 2 bp mismatches against the genomic DNA. Genome-editing efficiencies were assayed 6 days after transfection. Similarly, we transfected 15 Cas9-gRNAs with their corresponding ssODNs individually into PGP1-hiPSCs to target the same 15 sites and analyzed the efficiency 6 days after transfection. Bottom: the genome-editing efficiency of re-TALENs and Cas9-gRNAs targeting CCR5 in PGP1 hiPSCs. Panels 1 and 2 indicate NHEJ and HDR efficiencies mediated by reTALENs. Panels 3 and 4 indicate NHEJ and HDR efficiencies mediated by Cas9-gRNAs. NHEJ rates were calculated by the frequency of genomic alleles carrying deletions or insertions at the targeting region; HDR rates were calculated by the frequency of genomic alleles carrying 2 bp mismatches. Panel 5, the DNaseI HS profile of a hiPSC cell line from ENCODE database (Duke DNase HS, iPS NIHi7 DS). Of note, the scales of different panels are different.

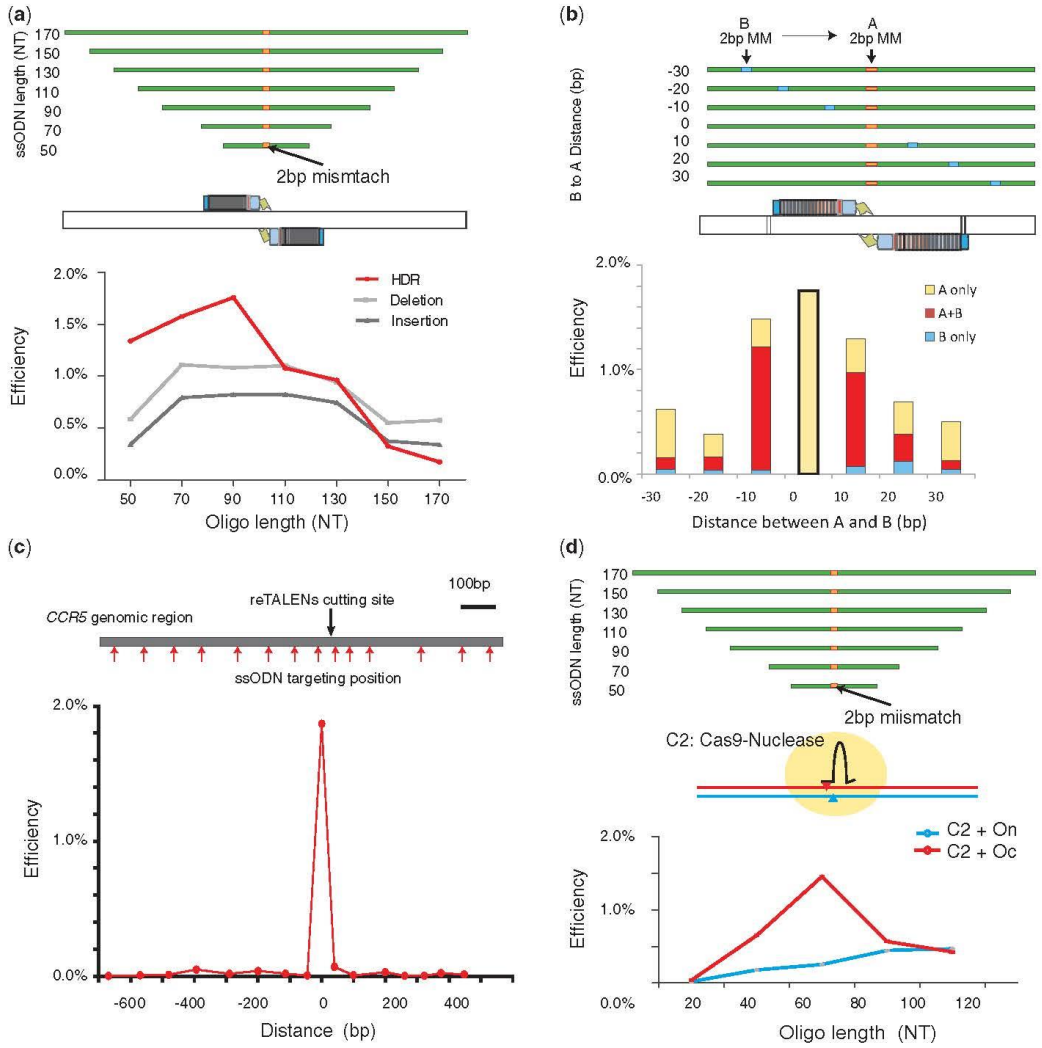


Figure 3. Study of functional parameters governing ssODN-mediated HDR with re-TALENs or Cas9-gRNAs in PGP1 hiPSCs. (**a**) PGP1 hiPSCs were co-transfected with re-TALENs pair (#3) and ssODNs of different lengths (50, 70, 90, 110, 130, 150 and 170 nt). All ssODNs possessed an identical 2 bp mismatch against the genomic DNA in the middle of their sequence. A 90 mer ssODN achieved optimal HDR in the targeted genome. The assessment of HDR, NHEJ-incurred deletion and insertion efficiency is described in the 'Materials and Methods' section. (**b**) 90 mer ssODNs corresponding to re-TALEN pair #3 each containing a 2 bp mismatch (**A**) in the center and an additional 2 bp mismatch (**B**) at different positions offset from A (where offsets varied from −30 to 30 bp) were used to test the effects of deviations from homology along the ssODN. Genome-editing efficiency of each ssODN was assessed in PGP1 hiPSCs. The bottom bar graph shows the incorporation frequency of A only, B only and A + B in the targeted genome. HDR rates decrease as the distance of homology deviations from the center increase (see text and Supplementary Figure S7a and b). (**c**) ssODNs targeted to sites with varying distances (−620∼480 bp) away from the target site of re-TALEN pair #3 were tested to assess the maximum distance within which we can place ssODNs to introduce mutations. All ssODNs carried a 2 bp mismatch in the middle of their sequences. We observed minimal HDR efficiency (≤0.06%) when the ssODN mismatch was positioned 40 bp away from the middle of re-TALEN pair's binding site. (**d**) PGP1 hiPSCs were co-transfected with Cas9-gRNA (AAVS1) and ssODNs of different orientation (O_c: complement to gRNA; O_n: non-complement to gRNA) and different lengths (30, 50, 70, 90 and 110 nt). All ssODNs possessed an identical 2 bp mismatch against the genomic DNA in the middle of their sequence. A 70 mer O_c achieved optimal HDR in the targeted genome.

still observe incorporation. A set of 90 nt ssODNs with central 2 bp mismatches targeting a range of larger distances (−600 to +400 bp) away from the re-TALEN-induced dsDNA break site were tested. When the ssODNs matched ≥40 bp away, we observed >30× lower HDR efficiencies compared with the control ssODN positioned centrally over the cut region (Figure 3c). The low level of incorporation that was observed may be due to processes unrelated to the dsDNA cut, as seen in experiments in which genomes are altered by a ssDNA donor alone (38). Meanwhile, the low level of HDR present when the ssODN is ~40 bp away may be due to a combination of weakened homology on the mismatch-containing side of the dsDNA cut along with insufficient ssODN oligo length on the other side of the dsDNA break.

We similarly tested the ssODNs DNA donor design for Cas9-gRNA-mediated targeting. First, we constructed Cas9-gRNA (C_2) targeting the AAVS1 locus and designed ssODN donors of variable orientations (O_c: complementary to the gRNA and O_n: non-complementary to the gRNA) and lengths (30, 50, 70, 90 and 110 nt). We found O_c achieved better efficiency than O_n, with a 70 mer O_c achieving an optimal HDR rate of 1.5%. (Figure 3d) The same ssODN strand bias was detected using a Cas9-derived nickase (C_c: Cas9_D10A), despite the fact that the HDR efficiencies mediated by C_c with ssODN were significantly less than C_2 (*t*-test, paired-end, $P = 0.02$) (Supplementary Figure S8). Future investigation will further elucidate the factors that may contribute to this bias, including sequence bias, direction of transcription and replication.

hiPSC clonal isolation of corrected cells

GEAS revealed that re-TALEN pair #3 achieved precise genome editing with an efficiency of ~1% in hiPSCs, a level at which correctly edited cells can usually be isolated by screening clones. HiPSCs have poor viability as single cells, but recent advances in culture conditions have facilitated outgrowth of hiPSCs from single cells (23). We optimized these protocols along with a single-cell FACS sorting procedure to establish a robust platform for single hiPSCs sorting and maintenance, where hiPSC clones can be recovered with survival rates of >25% (see 'Materials and Methods' section). We combined this method with a rapid and efficient genotyping system where we can conduct chromosomal DNA extraction and targeted genome amplification in 1-h single tube reactions, enabling large-scale genotyping of edited hiPSCs. Together, these methods comprise a pipeline for robustly obtaining genome-edited hiPSCs without selection.

To demonstrate this system (Figure 4a), we first transfected PGP1 hiPSCs with a pair of re-TALENs and an ssODN targeting CCR5 at site #3 (Supplementary Table S3), and we performed GEAS with a portion of the transfected cells, finding an HDR frequency of 1.7% (Figure 4b). This information, along with the 25% recovery of sorted single-cell clones, allowed us to estimate that we could obtain at least one correctly edited clone from five 96-well plates with Poisson probability 98% (assuming $\mu = 0.017 \times 0.25 \times 96 \times 5 \times 2$). Six days after transfection, hiPSCs were FACS sorted and 8 days after sorting, 100 hiPSC clones were screened. Sanger sequencing revealed that 2 of 100 of these unselected hiPSC colonies contained a heterozygous genotype possessing the 2 bp mutation introduced by the ssODN donor (Figure 4c). The targeting efficiency of 1% ($1\% = 2/2 \times 100$, 2 mono-allelic corrected clones out of 100 cell screened) was consistent with the next-generation sequencing analysis (1.7%) (Figure 4b). The pluripotency of the resulting hiPSCs was confirmed with immunostaining for SSEA4 and TRA-1-60 (Figure 4d). The successfully targeted hiPSCs clones were able to generate mature teratomas with features of all three germ layers (Figure 4e).

DISCUSSION

Here, we developed and demonstrated several improvements to the design and assessment of genome-editing reagents and demonstrated a streamlined method for efficient human stem cell editing. We first developed reTALENs, which simplify TALEN construction and enables the generation of functional lenti-viruses, which are important tools for delivering the reagents into many cell types and animals (33).

We then built a highly sensitive GEAS assay system to easily and precisely pinpoint and quantify HDR and NHEJ events in hiPSCs. In comparison with other methods of assessing design parameters for genome-editing, our genome-editing assessment tool provides simultaneous information on rates of HDR, NHEJ and other mutagenic processes through a single experimental and statistical analysis method versus performing different experiments and applying separate statistical methods for each individually. In the course of this study, we routinely pooled ~50 barcoded samples together and used the Illumina MiSeq system to obtain the sequence data, which was analyzed with our genome-editing assessment software. Currently, MiSeq can deliver ~20 Million paired-end 150 bp reads within 27 h so that up to 200 sample-barcoded targeting regions can be covered with ~100 K reads each at a cost of approximately $5 per sample. If desired, sample throughput can be traded off for higher sensitivity by allotting more reads per sample and processing fewer samples. Software and documentation for our genome-editing assessment system is available to provide researchers with the means to improve and standardize their genome-editing methods and extend them to additional cell lines and types.

Using our developed reTALENs, Cas9-gRNAs and GEAS method, we compared HDR and NHEJ efficiencies across 15 pairs of reTALENs and Cas9-gRNA (Supplementary Table S3 and S4) on the CCR5 locus. We found 13/15 of reTALEN pairs and all 15 Cas9-gRNAs exhibited detectable activities in hiPSCs, suggesting that both nuclease platforms serve as robust tools for genome editing. We confirmed the activity of the two failed reTALEN pairs in K562 cells and found 4 and 3% cutting efficiency, respectively, suggesting some

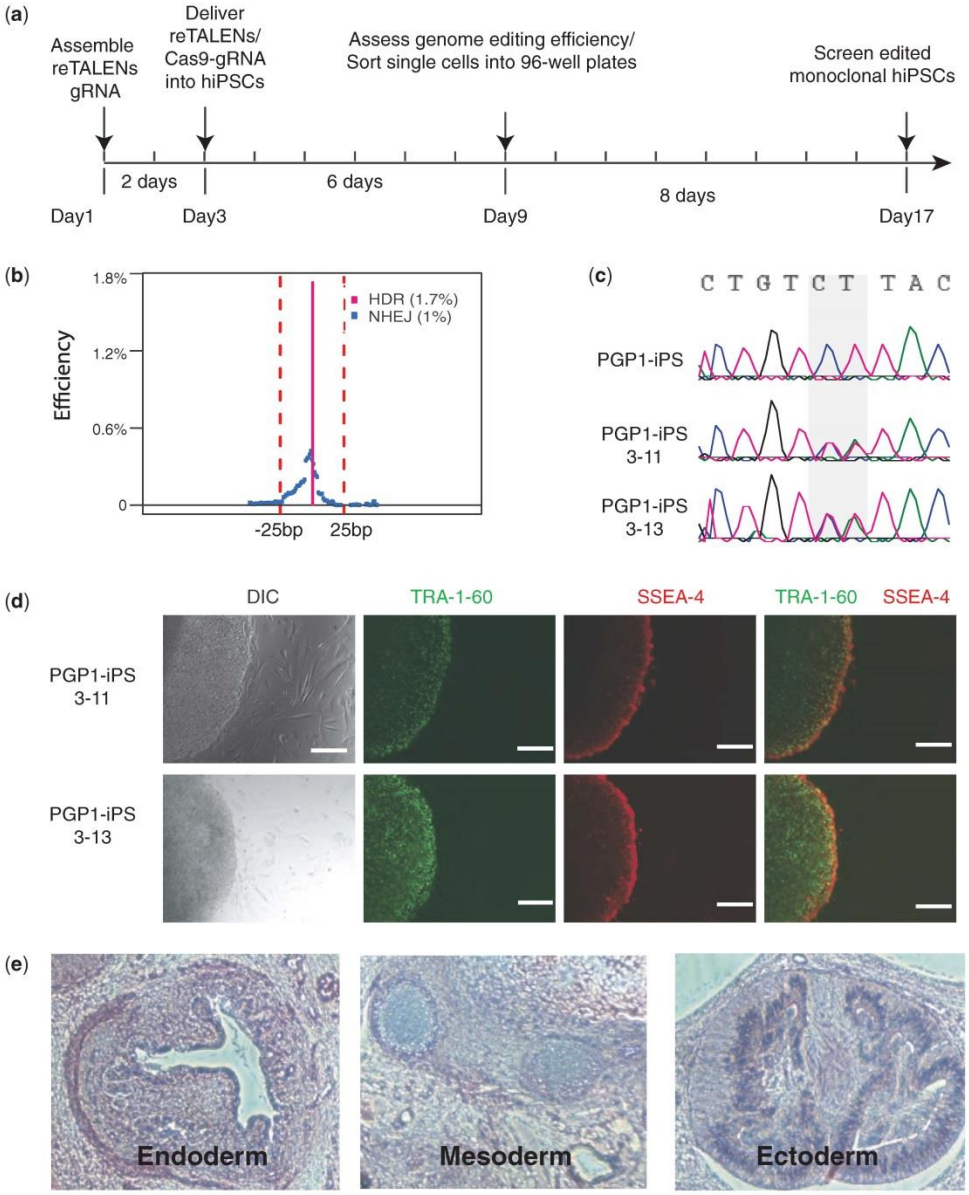


Figure 4. Using re-TALENs and ssODNs to obtain monoclonal genome-edited hiPSC without selection. **(a)** Timeline of the experiment. **(b)** Genome engineering efficiency of re-TALENs pair and ssODN (#3) assessed by the NGS platform described in Figure 2b. **(c)** Sanger sequencing results of monoclonal hiPSC colonies after genome editing. Of note, the 2 bp heterogeneous genotype (CT/CT→TA/CT) was successfully introduced into the genome of PGP1-iPS-3-11, PGP1-iPS-3-13 colonies. **(d)** Immunofluorescence staining of targeted PGP1-iPS-3-11. Cells were stained for the pluripotency markers Tra-1-60 and SSEA4. **(e)** Hematoxylin and eosin staining of teratoma sections generated from monoclonal PGP1-iPS-3-11 cells.

pertinent factors in hiPSCs, such as heterochromatin of methylation at the targeting regions make them resistant to reTALEN activity. In addition, we found that Cas9-gRNA induced on average 7 8× greater NHEJ rates than reTALEN, similar to recent reports (15). The effective concentration of Cas9-gRNA complexes or the intrinsic enzyme kinetics may contribute to this difference. Surprisingly, we did not see an equivalent increase of HDR with Cas9-gRNA and ssODN. Although ssODN concentration may reach saturating levels during construct delivery, ssODN availability at the DSB might be the limiting factor for HDR. Future studies using Cas9-gRNA nickases to generate defined DSB resections more favorable for HDR (36) can be conducted to test this hypothesis and further increase HDR efficiencies. Although we have compared the genome-targeting efficiencies achieved by reTALENs and Cas9-gRNA, a critical issue will also be to determine the generation of off-target mutations. It will be imperative to address the specificity of both targeting tools to improve the potential of hiPSCs genome engineering.

Finally, we demonstrated a streamlined pipeline for obtaining scarlessly edited human stem cells using our reagents. The pipeline comprises of the following: (i) reTALEN or Cas9-gRNA synthesis; (ii) prospective screening of reagents using GEAS; and (iii) high-throughput isolation of hiPSC clones. We note that with 1% HDR efficiency, it is feasible to generate isogeneic hiPSCs with mono-allelic mutations, which will facilitate hiPSC-based modeling of dominant alleles, allele-specific expression or X-linked mutations. However, targeting efficiencies must be improved to generate of homozygous mutations in hiPSCs. Other strategies such as transfection enrichment (15,17), or transient hypothermia (40), can be used together with our tools to achieve this goal. Last, we emphasize the versatility of our tools in that re-TALEs/Cas-gRNA can be engineered and used for other genomic-targeting technologies such as customized transcriptional factors and epigenetic modifiers, whereas GEAS can be applied to other gene-editing techniques, such as ZFNs, targeted nickases and meganucleases. We envision that our pipeline of efficiently generating scarlessly engineered human stem cells will allow the research community to resolve the causal underpinnings of numerous important biological problems, as well as to precisely engineer hiPSCs and other cell lines for autologous cell therapy.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank all the Church laboratory members for suggestion and support; and Daniel Gibson (J. Craig Venter Institute) for providing advice on assembly reactions.

FUNDING

National Human Genome Research Institute (NHGRI) Center for Excellence in Genomics Science [P50 HG005550, G.M.C.]; funded by Human Frontiers Science Program long-term fellowship (to M.G.). Funding for open access charge: NHGRI Center for Excellence in Genomics Science [P50 HG005550, G.M.C.].

Conflict of interest statement. G.M.C., L.Y, M.G. and J.Y. are inventors on a patent application describing the reTALE concept and assembly method.

REFERENCES

1. Carroll,D. (2011) Genome engineering with zinc-finger nucleases. *Genetics*, **188**, 773–782.
2. Wood,A.J., Lo,T.W., Zeitler,B., Pickle,C.S., Ralston,E.J., Lee,A.H., Amora,R., Miller,J.C., Leung,E., Meng,X. *et al.* (2011) Targeted genome editing across species using ZFNs and TALENs. *Science*, **333**, 307.
3. Perez-Pinera,P., Ousterout,D.G. and Gersbach,C.A. (2012) Advances in targeted genome editing. *Curr. Opin. Chem. Biol.*, **16**, 268–277.
4. Symington,L.S. and Gautier,J. (2011) Double-strand break end resection and repair pathway choice. *Annu. Rev. Genet.*, **45**, 247–271.
5. Urnov,F.D., Miller,J.C., Lee,Y.-L., Beausejour,C.M., Rock,J.M., Augustus,S., Jamieson,A.C., Porteus,M.H., Gregory,P.D. and Holmes,M.C. (2005) Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature*, **435**, 646–651.
6. Boch,J., Scholze,H., Schornack,S., Landgraf,A., Hahn,S., Kay,S., Lahaye,T., Nickstadt,A. and Bonas,U. (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, **326**, 1509–1512.
7. Cermak,T., Doyle,E.L., Christian,M., Wang,L., Zhang,Y., Schmidt,C., Baller,J.A., Somia,N.V., Bogdanove,A.J. and Voytas,D.F. (2011) Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.*, **39**, e82.
8. Briggs,A.W., Rios,X., Chari,R., Yang,L., Zhang,F., Mali,P. and Church,G.M. (2012) Iterative capped assembly: rapid and scalable synthesis of repeat-module DNA such as TAL effectors from individual monomers. *Nucleic Acids Res.*, **40**, e117.
9. Zhang,F., Cong,L., Lodato,S., Kosuri,S., Church,G.M. and Arlotta,P. (2011) LETTErs Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nat. Biotechnol.*, **29**, 149–154.
10. Reyon,D., Tsai,S.Q., Khayter,C., Foden,J.A., Sander,J.D. and Joung,J.K. (2012) FLASH assembly of TALENs for high-throughput genome editing. *Nat. Biotechnol.*, **30**, 460–465.
11. Holkers,M., Maggio,I., Liu,J., Janssen,J.M., Miselli,F., Mussolino,C., Recchia,A., Cathomen,T. and Gonçalves,M.A. (2012) Differential integrity of TALE nuclease genes following adenoviral and lentiviral vector gene transfer into human cells. *Nucleic Acids Res.*, **41**, e63.
12. Qiu,P., Shandilya,H., D'Alessio,J.M., O'Connor,K., Durocher,J. and Gerard,G.F. (2004) Mutation detection using Surveyor nuclease. *Biotechniques*, **36**, 702–707.
13. Miller,J.C., Tan,S., Qiao,G., Barlow,K.A., Wang,J., Xia,D.F., Meng,X., Paschon,D.E., Leung,E., Hinkley,S.J. *et al.* (2011) A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.*, **29**, 143–148.
14. Hockemeyer,D., Wang,H., Kiani,S., Lai,C.S., Gao,Q., Cassady,J.P., Cost,G.J., Zhang,L., Santiago,Y., Miller,J.C. *et al.* (2011) Genetic engineering of human pluripotent cells using TALE nucleases. *Nat. Biotechnol.*, **29**, 731–734.
15. Ding,Q., Lee,Y., Schaefer,E.A.K., Peters,D.T., Veres,A., Kim,K., Kuperwasser,N., Motola,D.L., Meissner,T.B., Hendriks,W.T. *et al.* (2013) Resource A TALEN genome-editing system for

generating human stem cell-based disease models. *Cell Stem Cell*, **12**, 238–251.
16. Mali,P., Yang,L., Esvelt,K.M., Aach,J., Guell,M., DiCarlo,J.E., Norville,J.E. and Church,G.M. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823–826.
17. Ding,Q., Regan,S.N., Xia,Y., Oostrom,L.A., Cowan,C.A. and Musunuru,K. (2013) Enhanced efficiency of human pluripotent stem cell genome editing through replacing TALENs with CRISPRs. *Cell Stem Cell*, **12**, 393–394.
18. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.
19. Cho,S.W., Kim,S., Kim,J.M. and Kim,J.S. (2013) Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.*, **31**, 230–232.
20. Hwang,W.Y., Fu,Y., Reyon,D., Maeder,M.L., Tsai,S.Q., Sander,J.D., Peterson,R.T., Yeh,J.-R.J. and Joung,J.K. (2013) Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat. Biotechnol.*, **31**, 227–229.
21. Chen,F., Pruett-Miller,S.M., Huang,Y., Gjoka,M., Duda,K., Taunton,J., Collingwood,T.N., Frodin,M. and Davis,G.D. (2011) High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nat. Methods*, **8**, 753–755.
22. Soldner,F., Laganière,J., Cheng,A.W., Hockemeyer,D., Gao,Q., Alagappan,R., Khurana,V., Golbe,L.I., Myers,R.H., Lindquist,S. *et al.* (2011) Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. *Cell*, **146**, 318–331.
23. Valamehr,B., Abujarour,R., Robinson,M., Le,T., Robbins,D., Shoemaker,D. and Flynn,P. (2012) A novel platform to enable the high-throughput derivation and characterization of feeder-free human iPSCs. *Sci. Rep.*, **2**, 213.
24. Sanjana,N.E., Cong,L., Zhou,Y., Cunniff,M.M., Feng,G. and Zhang,F. (2012) A transcription activator-like effector toolbox for genome engineering. *Nat. Protoc.*, **7**, 171–192.
25. Gibson,D.G., Young,L., Chuang,R., Venter,J.C., Iii,C.A.H., Smith,H.O. and America,N. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, **6**, 12–16.
26. Zou,J., Maeder,M.L., Mali,P., Pruett-Miller,S.M., Thibodeau-Beganny,S., Chou,B.K., Chen,G., Ye,Z., Park,I.H., Daley,G.Q. *et al.* (2009) Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells. *Cell Stem Cell*, **5**, 97–110.
27. Park,I.H., Lerou,P.H., Zhao,R., Huo,H. and Daley,G.Q. (2008) Generation of human-induced pluripotent stem cells. *Nat. Protoc.*, **3**, 1180–1186.
28. Hockemeyer,D., Wang,H., Kiani,S., Lai,C.S., Gao,Q., Cassady,J.P., Cost,G.J., Zhang,L., Santiago,Y., Miller,J.C. *et al.* (2011) Genetic engineering of human pluripotent cells using TALE nucleases. *Nature biotechnology*, **29**, 731–734.
29. Mussolino,C., Morbitzer,R., Lütge,F., Dannemann,N., Lahaye,T. and Cathomen,T. (2011) A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic Acids Res.*, **39**, 9283–9293.
30. Bedell,V.M., Wang,Y., Campbell,J.M., Poshusta,T.L., Starker,C.G., Krug Ii,R.G., Tan,W., Penheiter,S.G., Ma,A.C., Leung,A.Y.H. *et al.* (2012) *In vivo* genome editing using a high-efficiency TALEN system. *Nature*, **490**, 114–118.
31. Pathak,V.K. and Temin,H.M. (1990) Broad spectrum of in vivo forward mutations, hypermutations, and mutational hotspots in a retroviral shuttle vector after a single replication cycle: substitutions, frameshifts, and hypermutations. *Proc. Natl Acad. USA*, **87**, 6019–6023.
32. Tian,J., Ma,K. and Saaem,I. (2009) Advancing high-throughput gene synthesis technology. *Mol. Biosyst.*, **5**, 714–722.
33. Zou,J., Mali,P., Huang,X., Dowey,S.N. and Cheng,L. (2011) Site-specific gene correction of a point mutation in human iPS cells derived from an adult patient with sickle cell disease. *Blood*, **118**, 4599–4608.
34. Mali,P., Yang,L., Esvelt,K.M., Aach,J., Guell,M., Dicarlo,J.E., Norville,J.E. and Church,G.M. (2013) RNA-guided human genome engineering via Cas9. *Science (New York, N.Y.)*, **339**, 823–826.
35. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
36. Orlando,S.J., Santiago,Y., DeKelver,R.C., Freyvert,Y., Boydston,E.A., Moehle,E.A., Choi,V.M., Gopalan,S.M., Lou,J.F., Li,J. *et al.* (2010) Zinc-finger nuclease-driven targeted integration into mammalian genomes using donors with limited chromosomal homology. *Nucleic Acids Res.*, **38**, e152.
37. Wang,Z., Zhou,Z.J., Liu,D.P. and Huang,J.D. (2008) Double-stranded break can be repaired by single-stranded oligonucleotides via the ATM/ATR pathway in mammalian cells. *Oligonucleotides*, **18**, 21–32.
38. Rios,X., Briggs,A.W., Christodoulou,D., Gorham,J.M., Seidman,J.G. and Church,G.M. (2012) Stable gene targeting in human cells using single-strand oligonucleotides with modified bases. *PloS One*, **7**, e36697.
39. Elliott,B., Richardson,C., Winderbaum,J., Jac,A., Jasin,M. and Nickoloff,J.A.C.A. (1998) Gene conversion tracts from double-strand break repair in mammalian cells. *Mol. Cell. Biol.*, **18**, 93–101.
40. Doyon,Y., Choi,V.M., Xia,D.F., Vo,T.D., Gregory,P.D. and Holmes,M.C. (2010) Transient cold shock enhances zinc-finger nuclease-mediated gene disruption. *Nat. Methods*, **7**, 459–460.

Appendix B

RNA-Guided Human Genome Engineering via Cas9

RNA-Guided Human Genome Engineering via Cas9
Prashant Mali *et al.*
Science **339**, 823 (2013);
DOI: 10.1126/science.1232033

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

4. A. J. Wood *et al.*, *Science* **333**, 307 (2011).
5. M. Christian *et al.*, *Genetics* **186**, 757 (2010).
6. F. Zhang *et al.*, *Nat. Biotechnol.* **29**, 149 (2011).
7. J. C. Miller *et al.*, *Nat. Biotechnol.* **29**, 143 (2011).
8. D. Reyon *et al.*, *Nat. Biotechnol.* **30**, 460 (2012).
9. J. Boch *et al.*, *Science* **326**, 1509 (2009).
10. M. J. Moscou, A. J. Bogdanove, *Science* **326**, 1501 (2009).
11. B. L. Stoddard, *Q. Rev. Biophys.* **38**, 49 (2005).
12. M. Jinek *et al.*, *Science* **337**, 816 (2012).
13. G. Gasiunas, R. Barrangou, P. Horvath, V. Siksnys, *Proc. Natl. Acad. Sci. U.S.A.* **109**, E2579 (2012).
14. J. E. Garneau *et al.*, *Nature* **468**, 67 (2010).
15. H. Deveau, J. E. Garneau, S. Moineau, *Annu. Rev. Microbiol.* **64**, 475 (2010).
16. P. Horvath, R. Barrangou, *Science* **327**, 167 (2010).
17. K. S. Makarova *et al.*, *Nat. Rev. Microbiol.* **9**, 467 (2011).
18. D. Bhaya, M. Davison, R. Barrangou, *Annu. Rev. Genet.* **45**, 273 (2011).
19. E. Deltcheva *et al.*, *Nature* **471**, 602 (2011).
20. R. Sapranauskas *et al.*, *Nucleic Acids Res.* **39**, 9275 (2011).
21. A. H. Magadán, M. E. Dupuis, M. Villion, S. Moineau, *PLoS ONE* **7**, e40913 (2012).
22. H. Deveau *et al.*, *J. Bacteriol.* **190**, 1390 (2008).
23. F. J. Mojica, C. Díez-Villaseñor, J. García-Martínez, C. Almendros, *Microbiology* **155**, 733 (2009).
24. M. Jinek, J. A. Doudna, *Nature* **457**, 405 (2009).
25. C. D. Malone, G. J. Hannon, *Cell* **136**, 656 (2009).
26. G. Meister, T. Tuschl, *Nature* **431**, 343 (2004).
27. M. T. Certo *et al.*, *Nat. Methods* **8**, 671 (2011).
28. P. Mali *et al.*, *Science* **339**, 823 (2013).
29. P. A. Carr, G. M. Church, *Nat. Biotechnol.* **27**, 1151 (2009).

Acknowledgments: We thank the entire Zhang lab for their support and advice; P. A. Sharp for generous help with Northern blot analysis; C. Jennings, R. Desimone, and M. Kowalczyk for helpful comments; and X. Ye for help with confocal imaging. L.C. and X.W. are Howard Hughes Medical Institute International Student Research Fellows. D.C. is supported by the Medical Scientist Training Program. P.D.H. is a James Mills Pierce Fellow. X.W. is supported by NIH grants R01-GM34277 and R01-CA133404 to P. A. Sharp, X.W.'s thesis adviser. L.A.M. is supported by Searle Scholars, R. Allen, an Irma T. Hirschl Award, and a NIH Director's New Innovator Award (DP2AI104556). F.Z. is supported by a NIH Director's Pioneer Award (DP1MH100706); the Keck, McKnight, Gates, Damon Runyon, Searle Scholars, Klingenstein, and Simons foundations; R. Metcalfe; M. Boylan; and J. Pauley. The authors have no conflicting financial interests. A patent application has been filed relating to this work, and the authors plan on making the reagents widely available to the academic community through Addgene and to provide software tools via the Zhang lab Web site (www.genome-engineering.org).

Supplementary Materials

www.sciencemag.org/cgi/content/full/science.1231143/DC1
Materials and Methods
Figs. S1 to S8
Tables S1 and S2
References (*30*–*32*)

5 October 2012; accepted 12 December 2012
Published online 3 January 2013;
10.1126/science.1231143

RNA-Guided Human Genome Engineering via Cas9

Prashant Mali,[1]* Luhan Yang,[1,3]* Kevin M. Esvelt,[2] John Aach,[1] Marc Guell,[1] James E. DiCarlo,[4] Julie E. Norville,[1] George M. Church[1,2]†

Bacteria and archaea have evolved adaptive immune defenses, termed clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) systems, that use short RNA to direct degradation of foreign nucleic acids. Here, we engineer the type II bacterial CRISPR system to function with custom guide RNA (gRNA) in human cells. For the endogenous AAVS1 locus, we obtained targeting rates of 10 to 25% in 293T cells, 13 to 8% in K562 cells, and 2 to 4% in induced pluripotent stem cells. We show that this process relies on CRISPR components; is sequence-specific; and, upon simultaneous introduction of multiple gRNAs, can effect multiplex editing of target loci. We also compute a genome-wide resource of ~190 K unique gRNAs targeting ~40.5% of human exons. Our results establish an RNA-guided editing tool for facile, robust, and multiplexable human genome engineering.

Bacterial and archaeal clustered regularly interspaced short palindromic repeats (CRISPR) systems rely on CRISPR RNAs (crRNAs) in complex with CRISPR-associated (Cas) proteins to direct degradation of complementary sequences present within invading viral and plasmid DNA (*1*–*3*). A recent in vitro reconstitution of the *Streptococcus pyogenes* type II CRISPR system demonstrated that crRNA fused to a normally trans-encoded tracrRNA is sufficient to direct Cas9 protein to sequence-specifically cleave target DNA sequences matching the crRNA (*4*). The fully defined nature of this two-component system suggested that it might function in the cells of eukaryotic organisms such as yeast, plants, and even mammals. By cleaving genomic sequences targeted by RNA sequences (*4*–*6*), such a system could greatly enhance the ease of genome engineering.

Here, we engineer the protein and RNA components of this bacterial type II CRISPR system in human cells. We began by synthesizing a human codon optimized version of the Cas9 protein bearing a C-terminal SV40 nuclear localization signal and cloning it into a mammalian expression system (Fig. 1A and fig. S1A). To direct Cas9 to cleave sequences of interest, we expressed crRNA-tracrRNA fusion transcripts, hereafter referred to as guide RNAs (gRNAs), from the human U6 polymerase III promoter. Directly transcribing gRNAs allowed us to avoid reconstituting the RNA-processing machinery used by bacterial CRISPR systems (Fig. 1A and fig. S1B) (*4*, *7*–*9*). Constrained only by U6 transcription initiating with G and the requirement for the PAM (protospacer-adjacent motif) sequence -NGG following the 20 base pair (bp) crRNA target, our highly versatile approach can, in principle, target any genomic site of the form $GN_{20}GG$ (fig. S1C; see supplementary text S1 for a detailed discussion).

To test the functionality of our implementation for genome engineering, we developed a green fluorescent protein (GFP) reporter assay (Fig. 1B) in human embryonic kidney HEK 293T cells similar to one previously described (*10*). Specifically, we established a stable cell line bearing a genomically integrated GFP coding sequence disrupted by the insertion of a stop codon and a 68-bp genomic fragment from the AAVS1 locus that renders the expressed protein fragment nonfluorescent. Homologous recombination (HR) using an appropriate repair donor can restore the normal GFP sequence, which enabled us to quantify the resulting GFP^+ cells by flow-activated cell sorting (FACS).

To test the efficiency of our system at stimulating HR, we constructed two gRNAs, T1 and T2, that target the intervening AAVS1 fragment (Fig. 1B) and compared their activity to that of a previously described TAL effector nuclease heterodimer (TALEN) targeting the same region (*11*). We observed successful HR events using all three targeting reagents, with gene correction rates using the T1 and T2 gRNAs approaching 3% and 8%, respectively (Fig. 1C). This RNA-mediated editing process was notably rapid, with the first detectable GFP^+ cells appearing ~20 hours post transfection compared with ~40 hours for the AAVS1 TALENs. We observed HR only upon simultaneous introduction of the repair donor, Cas9 protein, and gRNA, which confirmed that all components are required for genome editing (fig. S2). Although we noted no apparent toxicity associated with Cas9/gRNA expression, work with zinc finger nucleases (ZFNs) and TALENs has shown that nicking only one strand further reduces toxicity. Accordingly, we also tested a Cas9D10A mutant that is known to function as a nickase in vitro, which yielded similar HR but lower nonhomologous end joining (NHEJ) rates (fig. S3) (*4*, *5*). Consistent with (*4*), in which a related Cas9 protein is shown to cut both strands

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [2]Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, MA 02138, USA. [3]Biological and Biomedical Sciences Program, Harvard Medical School, Boston, MA 02115, USA. [4]Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: gchurch@genetics.med.harvard.edu

3 bp upstream of the PAM, our NHEJ data confirmed that most deletions or insertions occurred at the 3′ end of the target sequence (fig. S3B). We also confirmed that mutating the target genomic site prevents the gRNA from effecting HR at that locus, which demonstrates that CRISPR-mediated genome editing is sequence-specific (fig. S4). Finally, we showed that two gRNAs targeting sites in the GFP gene, and also three additional gRNAs targeting fragments from homologous regions of the DNA methyl transferase 3a (DNMT3a) and DNMT3b genes could sequence-specifically induce significant HR in the engineered reporter cell lines (figs. S5 and S6). Together, these results confirm that RNA-guided genome targeting in human cells is simple to execute and induces robust HR across multiple target sites.

Having successfully targeted an integrated reporter, we next turned to modifying a native locus. We used the gRNAs described above to target the AAVS1 locus located in the PPP1R12C gene on chromosome 19, which is ubiquitously expressed across most tissues (Fig. 2A). We targeted 293Ts, human chronic myelogenous leukemia K562 cells, and PGP1 human induced pluripotent stem (iPS) cells (*12*) and analyzed the results by next-generation sequencing of the targeted locus. Consistent with our results for the GFP reporter assay, we observed high numbers of NHEJ events at the endogenous locus for all three cell types. The two gRNAs T1 and T2 achieved NHEJ rates of 10 and 25% in 293Ts, 13 and 38% in K562s, and 2 and 4% in PGP1-iPS cells, respectively (Fig. 2B). We observed no overt toxicity from the Cas9 and gRNA expression required to induce NHEJ in any of these cell types. As expected, NHEJ-mediated deletions for T1 and T2 were centered around the target site positions, which further validated the sequence-specificity of this targeting process (figs. S7 to S9). Simultaneous introduction of both T1 and T2 gRNAs resulted in high-efficiency deletion of the intervening 19-bp fragment (fig. S8), which demonstrated that multiplexed editing of genomic loci is feasible using this approach.

Last, we attempted to use HR to integrate either a double-stranded DNA donor construct (*13*) or an oligo donor into the native AAVS1 locus (Fig. 2C and fig. S10). We confirmed HR-mediated integration, using both approaches, by polymerase chain reaction (PCR) (Fig. 2D and fig. S10) and Sanger sequencing (Fig. 2E). We also readily derived 293T or iPS clones from the pool of modified cells using puromycin selection over 2 weeks (Fig. 2F and fig. S10). These results demonstrate that this approach enables efficient integration of foreign DNA at endogenous loci in human cells.

Our versatile RNA-guided genome-editing system can be readily adapted to modify other genomic sites by simply modifying the sequence of our gRNA expression vector to match a compatible sequence in the locus of interest. To facilitate this process, we bioinformatically generated ~190,000 specific gRNA-targetable sequences

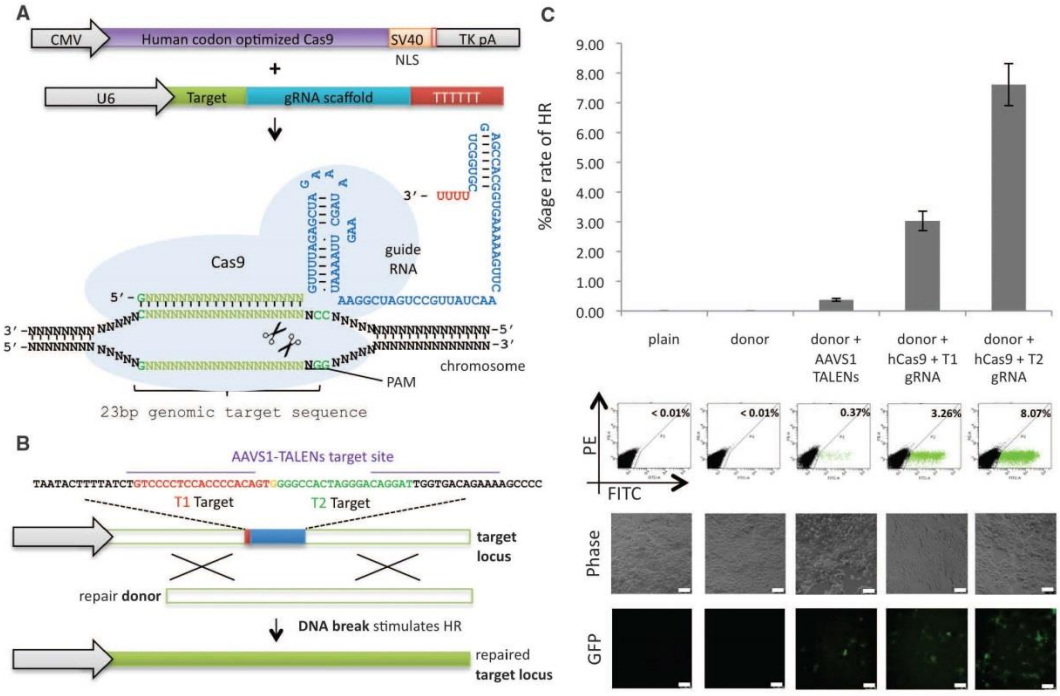


Fig. 1. Genome editing in human cells using an engineered type II CRISPR system. (**A**) RNA-guided gene targeting in human cells involves coexpression of the Cas9 protein bearing a C-terminal SV40 nuclear localization signal (NLS) with one or more gRNAs expressed from the human U6 polymerase III promoter. Cas9 unwinds the DNA duplex and cleaves both strands upon recognition of a target sequence by the gRNA, but only if the correct PAM is present at the 3′ end. Any genomic sequence of the form $GN_{20}GG$ can, in principle, be targeted. CMV, cytomegalovirus promoter; TK, thymidine kinase; pA, polyadenylation signal. (**B**) A genomically integrated GFP coding sequence is disrupted by the insertion of a stop codon and a 68-bp genomic fragment from the AAVS1 locus. Restoration of the GFP sequence by HR with an appropriate donor sequence results in GFP^+ cells that can be quantified by FACS. T1 and T2 gRNAs target sequences within the AAVS1 fragment. Binding sites for the two halves of the TALEN are underlined. (**C**) Bar graph depicting HR efficiencies induced by T1, T2, and TALEN-mediated nuclease activity at the target locus, as measured by FACS. Representative FACS plots and microscopy images of the targeted cells are depicted below. (Scale bar, 100 μm.) Data are shown as means ± SEM ($N = 3$).

targeting ~40.5% exons of genes in the human genome (refer to methods and table S1). We also incorporated these target sequences into a 200-bp format compatible with multiplex synthesis on DNA arrays (*14*) (fig. S11 and tables S2 and S3). This resource provides a ready genome-wide reference of potential target sites in the human genome and a methodology for multiplex gRNA synthesis.

Our results demonstrate the promise of CRISPR-mediated gene targeting for RNA-guided, robust, and multiplexable mammalian genome engineering. The ease of retargeting our system to modify genomic sequences greatly exceeds that of comparable ZFNs and TALENs, while offering similar or greater efficiencies (*4*). Existing studies of type II CRISPR specificity (*4*) suggest that target sites must perfectly match the PAM sequence NGG and the 8- to 12-base "seed sequence" at the 3′ end of the gRNA. The importance of the remaining 8 to 12 bases is less well understood and may depend on the binding strength of the matching gRNAs or on the inherent tolerance of Cas9 itself. Indeed, Cas9 will tolerate single mismatches at the 5′ end in bacteria and in vitro, which suggests that the 5′ G is not required. Moreover, it is likely that the target locus's underlying chromatin structure and epigenetic state will also affect the efficiency of genome editing in eukaryotic cells (*13*), although we suspect that Cas9's helicase activity may render it more robust to these factors, but this remains to be evaluated. Elucidating the frequency and underlying causes of off-target nuclease activity (*15*, *16*) induced by CRISPR, ZFN (*17*, *18*), and TALEN (*19*, *20*) genome-engineering

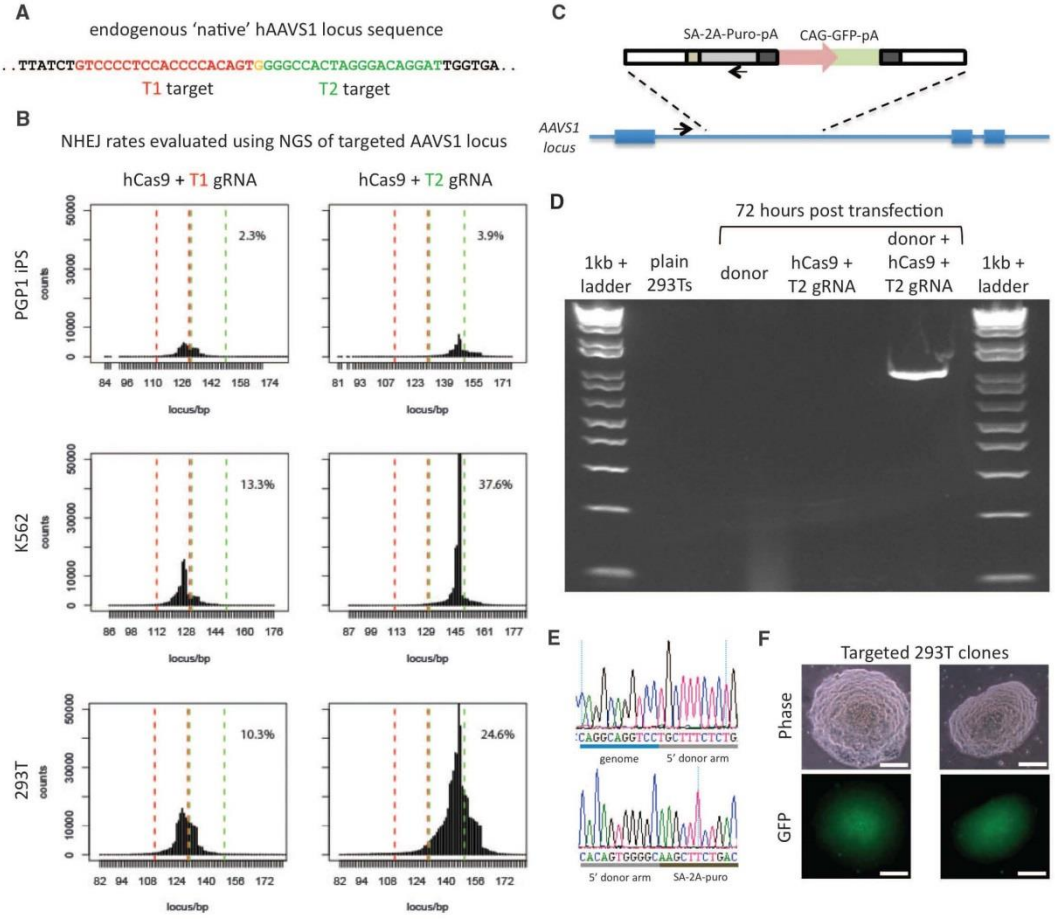


Fig. 2. RNA-guided genome editing of the native AAVS1 locus in multiple cell types. (**A**) T1 (red) and T2 (green) gRNAs target sequences in an intron of the PPP1R12C gene within the chromosome 19 AAVS1 locus. (**B**) Total count and location of deletions caused by NHEJ in 293Ts, K562s, and PGP1 iPS cells after expression of Cas9 and either T1 or T2 gRNAs as quantified by next-generation sequencing. Red and green dashed lines demarcate the boundaries of the T1 and T2 gRNA targeting sites. NHEJ frequencies for T1 and T2 gRNAs were 10% and 25% in 293T, 13% and 38% in K562, and 2% and 4% in PGP1 iPS cells, respectively. (**C**) DNA donor architecture for HR at the AAVS1 locus, and the locations of sequencing primers (arrows) for detecting successful targeted events, are depicted. (**D**) PCR assay 3 days after transfection demonstrates that only cells expressing the donor, Cas9 and T2 gRNA exhibit successful HR events. (**E**) Successful HR was confirmed by Sanger sequencing of the PCR amplicon, which showed that the expected DNA bases at both the genome-donor and donor-insert boundaries are present. (**F**) Successfully targeted clones of 293T cells were selected with puromycin for 2 weeks. Microscope images of two representative GFP+ clones is shown. (Scale bar, 100 μm.)

tools will be of utmost importance for safe genome modification and perhaps for gene therapy. Potential avenues for improving CRISPR specificity include evaluating Cas9 homologs identified through bioinformatics and directed evolution of these nucleases toward higher specificity. Similarly, the range of CRISPR-targetable sequences could be expanded through the use of homologs with different PAM requirements (*9*) or by directed evolution. Finally, inactivating one of the Cas9 nuclease domains increases the ratio of HR to NHEJ and may reduce toxicity (figs. S1A and fig. S3) (*4*, *5*), whereas inactivating both domains may enable Cas9 to function as a retargetable DNA binding protein. As we explore these areas, we note that another parallel study (*21*) has independently confirmed the high efficiency of CRISPR-mediated gene targeting in mammalian cell lines. We expect that RNA-guided genome targeting will have broad implications for synthetic biology (*22*, *23*), the direct and multiplexed perturbation of gene networks (*13*, *24*), and targeted ex vivo (*25*–*27*) and in vivo gene therapy (*28*).

References and Notes

1. B. Wiedenheft, S. H. Sternberg, J. A. Doudna, *Nature* **482**, 331 (2012).
2. D. Bhaya, M. Davison, R. Barrangou, *Annu. Rev. Genet.* **45**, 273 (2011).
3. M. P. Terns, R. M. Terns, *Curr. Opin. Microbiol.* **14**, 321 (2011).
4. M. Jinek *et al.*, *Science* **337**, 816 (2012).
5. G. Gasiunas, R. Barrangou, P. Horvath, V. Siksnys, *Proc. Natl. Acad. Sci. U.S.A.* **109**, E2579 (2012).
6. R. Sapranauskas *et al.*, *Nucleic Acids Res.* **39**, 9275 (2011).
7. T. R. Brummelkamp, R. Bernards, R. Agami, *Science* **296**, 550 (2002).
8. M. Miyagishi, K. Taira, *Nat. Biotechnol.* **20**, 497 (2002).
9. E. Deltcheva *et al.*, *Nature* **471**, 602 (2011).
10. J. Zou, P. Mali, X. Huang, S. N. Dowey, L. Cheng, *Blood* **118**, 4599 (2011).
11. N. E. Sanjana *et al.*, *Nat. Protoc.* **7**, 171 (2012).
12. J. H. Lee *et al.*, *PLoS Genet.* **5**, e1000718 (2009).
13. D. Hockemeyer *et al.*, *Nat. Biotechnol.* **27**, 851 (2009).
14. S. Kosuri *et al.*, *Nat. Biotechnol.* **28**, 1295 (2010).
15. V. Pattanayak, C. L. Ramirez, J. K. Joung, D. R. Liu, *Nat. Methods* **8**, 765 (2011).
16. N. M. King, O. Cohen-Haguenauer, *Mol. Ther.* **16**, 432 (2008).
17. Y. G. Kim, J. Cha, S. Chandrasegaran, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 1156 (1996).
18. E. J. Rebar, C. O. Pabo, *Science* **263**, 671 (1994).
19. J. Boch *et al.*, *Science* **326**, 1509 (2009).
20. M. J. Moscou, A. J. Bogdanove, *Science* **326**, 1501 (2009).
21. L. Cong *et al.*, *Science* **339**, 819 (2013).
22. A. S. Khalil, J. J. Collins, *Nat. Rev. Genet.* **11**, 367 (2010).
23. P. E. Purnick, R. Weiss, *Nat. Rev. Mol. Cell Biol.* **10**, 410 (2009).
24. J. Zou *et al.*, *Cell Stem Cell* **5**, 97 (2009).
25. N. Holt *et al.*, *Nat. Biotechnol.* **28**, 839 (2010).
26. F. D. Urnov *et al.*, *Nature* **435**, 646 (2005).
27. A. Lombardo *et al.*, *Nat. Biotechnol.* **25**, 1298 (2007).
28. H. Li *et al.*, *Nature* **475**, 217 (2011).

Acknowledgments: This work was supported by NIH grant P50 HG005550. We thank S. Kosuri for advice on the oligonucleotide pool designs and synthesis. G.M.C. and P.M. have filed a patent based on the findings of this study.

Supplementary Materials

www.sciencemag.org/cgi/content/full/science.1232033/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S11
Tables S1 to S3
References (*29*–*46*)

26 October 2012; accepted 12 December 2012
Published online 3 January 2013;
10.1126/science.1232033

Cyclic GMP-AMP Is an Endogenous Second Messenger in Innate Immune Signaling by Cytosolic DNA

Jiaxi Wu,[1]* Lijun Sun,[1,2]* Xiang Chen,[1] Fenghe Du,[1] Heping Shi,[3] Chuo Chen,[3] Zhijian J. Chen[1,2]†

Cytosolic DNA induces type I interferons and other cytokines that are important for antimicrobial defense but can also result in autoimmunity. This DNA signaling pathway requires the adaptor protein STING and the transcription factor IRF3, but the mechanism of DNA sensing is unclear. We found that mammalian cytosolic extracts synthesized cyclic guanosine monophosphate–adenosine monophosphate (cyclic GMP-AMP, or cGAMP) in vitro from adenosine triphosphate and guanosine triphosphate in the presence of DNA but not RNA. DNA transfection or DNA virus infection of mammalian cells also triggered cGAMP production. cGAMP bound to STING, leading to the activation of IRF3 and induction of interferon-β. Thus, cGAMP in metazoans and functions as an endogenous second messenger that triggers interferon production in response to cytosolic DNA.

Host defense against foreign genetic elements is one of the most fundamental functions of a living organism. The presence of self or foreign DNA in the cytoplasm is sensed by eukaryotic cells as a danger signal or a sign of foreign invasion (*1*). DNA can be introduced into the cytoplasm by bacterial or viral infection, transfection, or "leakage" from the nucleus or mitochondria under some pathological conditions that cause autoimmune diseases such as lupus. In mammalian cells, cytosolic DNA triggers the production of type I interferons and other cytokines through the endoplasmic reticulum protein STING (also known as MITA, MPYS, or ERIS) (*2*). STING recruits and activates the cytosolic kinases IKK and TBK1, which activate the transcription factors NF-κB and IRF3, respectively. NF-κB and IRF3 then enter the nucleus and function together to induce interferons and other cytokines. DNA-dependent RNA polymerase III has been shown to be a sensor that detects and transcribes AT-rich DNAs such as poly(deoxyadenosine-deoxythymidine) [poly(dA:dT)] into an RNA ligand capable of stimulating the RIG-I pathway to induce interferons (*3*, *4*). However, most DNA sequences do not activate the RNA polymerase III RIG-I pathway. Instead, cytosolic DNA activates the STING-dependent pathway in a sequence-independent manner. How cytosolic DNA activates the STING pathway remains elusive.

We hypothesized that DNA binds to and activates a putative cytosolic DNA sensor, which then directly or indirectly activates STING, leading to the activation of IRF3 and NF-κB (fig. S1A). To test this model, we developed an in vitro complementation assay using the murine fibrosarcoma cell line L929, which is known to induce interferon-β (IFN-β) in a STING-dependent manner (*5*) (Fig. 1A). We used an L929 cell line stably expressing a short hairpin RNA (shRNA) against STING such that DNA transfection would only activate factors upstream of STING, including the putative DNA sensor (fig. S1, A and B). The L929-shSTING cells were transfected with different types of DNA, and then cytoplasmic extracts from these cells were mixed with the human monocytic cell line THP1 or murine macrophage cell line Raw264.7, which was permeabilized with perfringolysin O (PFO; Fig. 1A). PFO treatment pokes holes in the plasma membrane (*6*), allowing the cytoplasm to diffuse in and out of cells, while retaining organelles including the endoplasmic reticulum (which contains STING) and the Golgi apparatus inside the cells (*7*). If an upstream activator of STING is generated in the DNA-transfected cells, the cytoplasm containing such an activator is expected to activate STING in the PFO-permeabilized cells, leading to the phosphorylation and dimerization of IRF3.

Cytoplasmic extracts from L929-shSTING cells transfected with a DNA sequence known as interferon-stimulatory DNA (ISD; Fig. 1B, lane 2), poly(dA:dT), a GC-rich 50 base pair

[1]Department of Molecular Biology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. [2]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. [3]Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: zhijian.chen@utsouthwestern.edu

Appendix C

Patent Application of Targeted deaminases

US 20110104787A1

(19) **United States**

(12) **Patent Application Publication**

Church et al.

(10) **Pub. No.: US 2011/0104787 A1**

(43) **Pub. Date: May 5, 2011**

(54) **FUSION PEPTIDES THAT BIND TO AND MODIFY TARGET NUCLEIC ACID SEQUENCES**

(75) Inventors: **George M. Church**, Brookline, MA (US); **Luhan Yang**, Cambridge, MA (US)

(73) Assignee: **President and Fellows of Harvard College**, Cambridge, MA (US)

(21) Appl. No.: **12/939,505**

(22) Filed: **Nov. 4, 2010**

Related U.S. Application Data

(60) Provisional application No. 61/258,336, filed on Nov. 5, 2009.

Publication Classification

(51) **Int. Cl.**

C12N 9/78	(2006.01)
C07H 21/00	(2006.01)
C12N 15/63	(2006.01)
C12N 5/10	(2006.01)
C12N 5/071	(2010.01)

(52) **U.S. Cl.** **435/227**; 536/23.2; 435/320.1; 435/325; 435/366; 435/375

(57) **ABSTRACT**

Novel methods and compositions for altering target nucleic acid (e.g., DNA e.g., genomic DNA) sequences are provided. Fusion proteins including one or more DNA binding domains and one or more DNA modifying domains are provided. Isolated polynucleotides encoding fusion proteins including one or more DNA binding domains and one or more DNA modifying domains are provided.

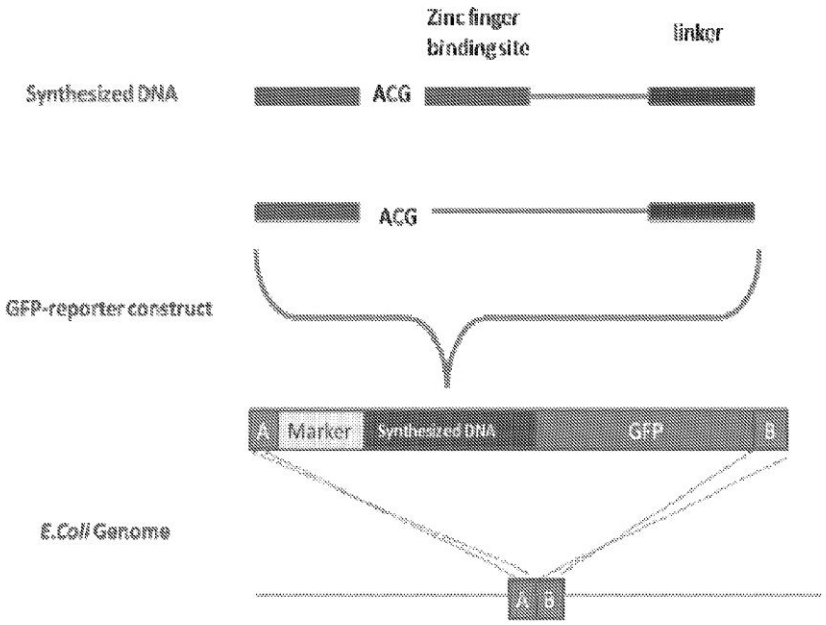

Zinc finger binding site
linker
Synthesized DNA
ACG
ACG
GFP-reporter construct
A
Marker
Synthesized DNA
GFP
B
E.Coli Genome
A
B

Figure 1

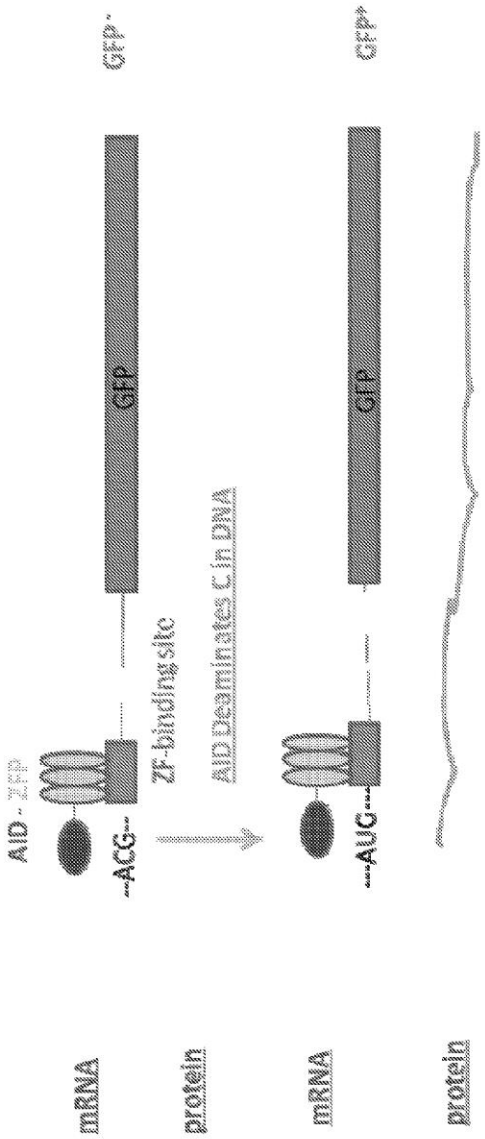

GFP-
GFP
ZF-binding site
AID Deaminates C in DNA
GFP*
GFP
AID - ZFP
---ACG---
---AUG---
mRNA
protein
mRNA
protein

Figure 2

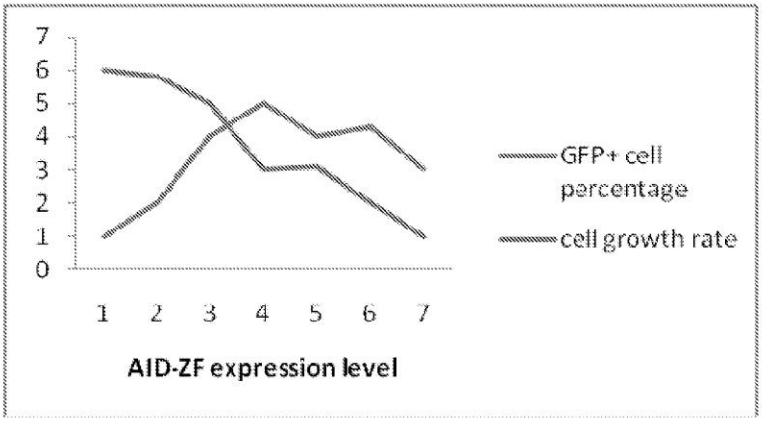

7
6
5
4
3
2
1
0
1 2 3 4 5 6 7
AID-ZF expression level
GFP+ cell percentage
cell growth rate

Figure 3

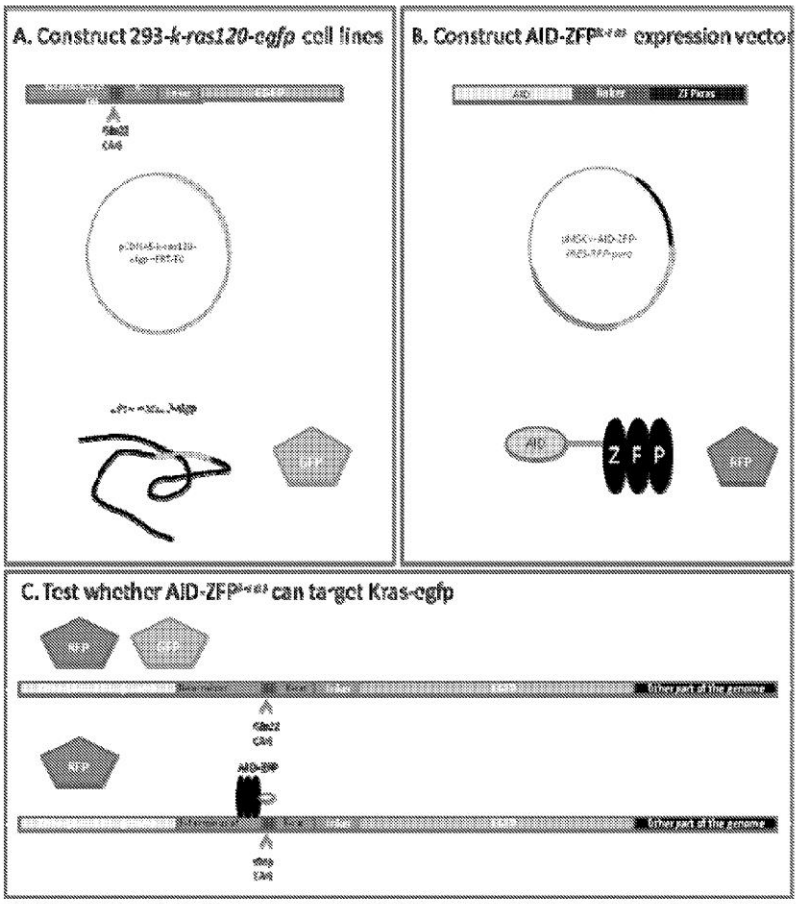
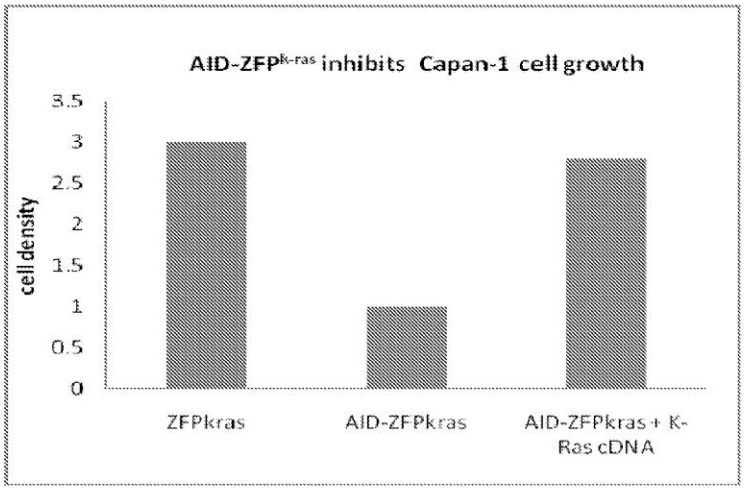

A. Construct 293-k-ras120-egfp cell lines
B. Construct AID-ZFP
expression vecto
AID
Linker
ZF Parts
Z F P
C. Test whether AID-ZFP can target Kras-egfp

Figure 4


AID-ZFP$^{k\text{-}ras}$ inhibits Capan-1 cell growth
cell density
3.5
3
2.5
2
1.5
1
0.5
0
ZFPkras
AID-ZFPkras
AID-ZFPkras + K-Ras cDNA

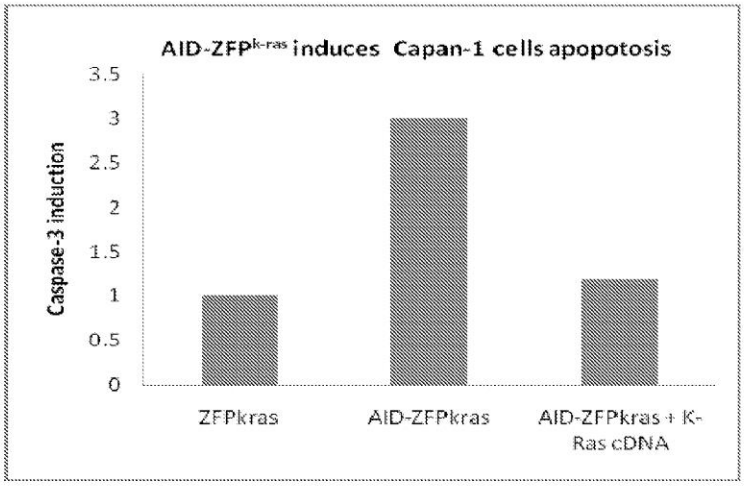

AID-ZFP$^{k\text{-}ras}$ induces Capan-1 cells apopotosis
Caspase-3 induction
3.5
3
2.5
2
1.5
1
0.5
0
ZFPkras
AID-ZFPkras
AID-ZFPkras + K-Ras cDNA

Figure 5

FUSION PEPTIDES THAT BIND TO AND MODIFY TARGET NUCLEIC ACID SEQUENCES

RELATED APPLICATIONS

[0001] This application claims priority to U.S. Provisional Patent Application No. 61/258,336, filed on Nov. 5, 2009 and is hereby incorporated herein by reference in its entirety for all purposes.

FIELD

[0002] Embodiments of the present invention relate in general to methods and compositions for altering target nucleic acid (e.g., genomic DNA) sequences.

BACKGROUND

[0003] Inducing multiple targeted mutations requires high efficiencies. Methods known in the art for inducing multiple targeted mutations include the use of single-stranded oligomers in strains with mismatch repair deficiency and expression of homologous DNA pairing proteins (e.g., lambda beta), or the use of nucleases and recombinases (e.g., Zn-finger nucelases, meanucleases, phage integrases and other microbial recombinases). Each of these methods shares the disadvantages of requiring three molecules to be simultaneously present (DNA donor, acceptor and protein catalyst) and most of them also can provoke DNA damage which does not repair in the desired manner.

SUMMARY

[0004] Methods and compositions for providing fusion proteins that functionally link one or more binding domains (e.g., DNA binding domains) with one or more modification domains (e.g., DNA modification domains) that alter one or more nucleosides of a target nucleic acid sequence (e.g., a target DNA sequence, such as, e.g., genomic DNA) are provided. The methods and compositions described herein provide advantages over current methods known in the art in that in contrast to art-known methods, no donor DNA needs to be coordinated in vivo with the action of the fusion proteins described herein.

[0005] The methods and compositions described herein address the need for the ability to engineer large numbers of sites in genomes, a need that is greatly increasing due to the growth of hypotheses based on dramatic increase in genomic sequence data. The methods and compositions described herein enable targeted homologous allele replacement, an approach to gene therapy that overcomes the limitations of relatively more random transfection or viral delivery which can result in unstable constructs and/or integration events which can induce cancer. The methods and compositions described herein will facilitate metabolic engineering (Wang et al. (2009) *Nature* 460(7257):894).

[0006] Accordingly, in certain exemplary embodiments, a non-naturally occurring fusion protein comprising a DNA binding domain, and a DNA modifying domain that includes a functional fragment of a deaminase protein (e.g., activation-induced deaminase (AID)), wherein the fusion protein is capable of binding to and altering a target oligonucleotide sequence (e.g., DNA (e.g., genomic DNA)) is provided. In certain aspects, the DNA binding domain includes one or more motifs selected from the group consisting of helix-turn-helix, leucine zipper, winged helix, winged helix turn helix, helix-loop-helix, zinc finger, immunoglobulin fold, B3 domain and TATA-box binding protein domain. In other aspects, an isolated polynucleotide (e.g., an expression vector) is provided that encodes the fusion protein. In certain aspects, the protein and/or isolated polynucleotide are present in a host cell.

[0007] In certain exemplary embodiments, a cell comprising a non-naturally occurring fusion protein, wherein the fusion protein includes a DNA binding domain, and a DNA modifying domain that includes a functional fragment of a deaminase protein (e.g., AID), wherein the fusion protein is capable of binding to and altering a target oligonucleotide sequence is provided. In certain aspects, the cell is an animal cell (e.g., a mammalian (e.g., human) cell). In other aspects, the cell is a stem cell (e.g., a hematopoietic stem cell).

[0008] In certain exemplary embodiments, a method of modulating expression of an endogenous gene in a cell is provided. In other exemplary embodiments, a method of inserting one or more exogenous nucleotide sequences and/or genes into a genome in a cell is provided. The method includes the steps of contacting a cell with a non-naturally occurring fusion protein wherein the fusion protein includes a DNA binding domain, and a DNA modifying domain including a functional fragment of a deaminase protein (e.g., AID), wherein the fusion protein is capable of binding to and altering an oligonucleotide sequence of an endogenous gene, and allowing the fusion protein to bind to and alter the oligonucleotide sequence of the endogenous gene to modulate expression of the endogenous gene. In certain aspects, all or part of an endogenous gene is excised from the genome. In certain aspects, the cell is an animal cell (e.g., a mammalian (e.g., human) cell). In other aspects, the cell is a stem cell (e.g., a hematopoietic stem cell). In certain aspects, expression of the endogenous gene is repressed. In other aspects, expression of the endogenous gene is activated.

[0009] Further features and advantages of certain embodiments of the present invention will become more fully apparent in the following description of the embodiments and drawings thereof, and from the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee. The foregoing and other features and advantages of the present invention will be more fully understood from the following detailed description of illustrative embodiments taken in conjunction with the accompanying drawings in which:

[0011] FIG. **1** schematically depicts the construction of a green fluorescent protein (GFP) assay platform. (1) a DNA fragment with a broken start codon "ACG," zinc finger (or other DNA binding domain) binding site and a linker coding region will be synthesized. (2) The GFP reporter construct has 50 bases of homology to the intended target (A, B blue cassette), a drug-resistance marker (yellow cassette), synthesized DNA (red cassette) harboring transcription and translation cis-elements, a broken start codon "ACG" and a linker coding region; a promoter-less GFP (green cassette). (3) This construct will be incorporated into the genome by recombination-mediated genetic engineering (recombineering).

[0012] FIG. **2** schematically depicts in vivo testing of the activity of AID-ZFP^{oct4}. Bacteria will show a GFP^{+} phenotype when the cytidine in the broken "ACG" start codon is mutated to T by activation-induced deaminase (AID)-mediated reaction.

[0013] FIG. **3** graphically depicts how the GFP^{+} cell percentage is expected to change with the expression level of AID-zinc finger protein (ZFP) fusion construct.

[0014] FIG. **4**A-**4**C schematically depict whether AID-ZFP$^{K\text{-}ras120}$ can specifically mutate K-ras Gln22(CAG) to a premature stop codon (TAG). (A) the K-ras120-egfp will be integrated into the 293 cell genome. 120 base pair K-ras120 gene fragment will be fused with egfp by a linker. The codon of Gln22, CAG, is shown in red. (B) 293-K-ras120-egfp cell will be transfected with AID-ZFP$^{K\text{-}ras120}$, and the success of transfection will be verified by RFP, which is co-translated with AID-ZFP$^{K\text{-}ras120}$. (C) If K-ras120-egfp is not targeted by AID-ZFP$^{K\text{-}ras}$, yellow fluorescence will be detected (RFP^{+} GFP). If K-ras120-egfp is targeted by AID-ZFP$^{K\text{-}ras120}$, which introduces a premature stop codon, only red fluorescence will be detected.

[0015] FIG. **5** graphically depicts that AID-ZFP$^{k\text{-}ras}$ will inhibit capan-1 cell growth and triggers apoptosis. In these experiments, ZFP$^{k\text{-}ras}$ is the negative control. To test whether AID-ZFP$^{k\text{-}ras}$-induced effects are the specific results of k-ras targeting, a K-Ras cDNA that loses the binding site of ZFP$^{k\text{-}ras}$ will be introduced into the cell to test whether it can rescue the phenotypes.

DETAILED DESCRIPTION

[0016] The fusion proteins described herein may be applied with particular advantage to modify target oligonucleotide (e.g., DNA) sequences. The methods and compositions described herein are particularly useful for targeted editing of genomic DNA as well as for genetically engineering cells (e.g., stem cells and the like).

[0017] In certain exemplary embodiments, polypeptides (e.g., fusion proteins) that are capable of interacting with and/or modifying a target nucleic acid (e.g., DNA) sequence are provided. As used herein, a "fusion" polypeptide refers to a polypeptide in which two or more subunit molecules are linked, e.g., covalently. The term "functionally linked," when describing the relationship between two polypeptides present as part of a fusion protein, refers to a juxtaposition wherein the regions are in a relationship permitting them to function in their intended manner. For example, a DNA binding domain "functionally linked" to a DNA modifying domain is ligated in such a way that one or more target nucleosides (e.g., of a target DNA) are enzymatically modified by the DNA modifying domain when the DNA binding domain is bound to the target oligonucleotide (e.g., DNA) sequence.

[0018] As used herein, the term "DNA binding domain" is intended to refer to, but is not limited to, a motif that can bind to a specific DNA sequence (e.g., a genomic DNA sequence). DNA binding domains have at least one motif that recognizes and binds to single-stranded or double-stranded DNA. DNA binding domains can interact with DNA in a sequence-specific (e.g., transcription factors, restriction enzymes, telomerase and the like) or a non-sequence-specific (e.g., *Drosophila melanogaster* HMG-D protein) manner. DNA binding domains can bind DNA at one or more of the major groove, the minor groove, and the sugar phosphate backbone. Proteins having DNA binding domains are well known in the art and include, but are not limited to, transcription factors, nucleases and structural proteins and the like and play roles in the replication, repair, storage, modification and expression of DNA. In certain exemplary embodiments, DNA binding domains from one or more DNA binding proteins are provided.

[0019] DNA binding domain motifs include, but are not limited to, the helix-turn-helix, the leucine zipper or bZIP, the winged helix, the winged helix turn helix, the helix-loop-helix, the zinc finger, the immunoglobulin fold, the B3 domain and the TBP-binding domain. For reviews of DNA binding domains and protein structure motifs, See Branden and Tooze (1991) *Protein Structure and Function*, Garland Pub.; Voet, Voet and Pratt (2001) *Fundamentals of Biochemistry*, Ch. 23, Wiley Pub.; Stryer (1995) *Biochemistry* (4th ed.), Ch. 33, 36, 37, W.H. Freeman & Company; Lehninger (2004) *Principles of Biochemistry* (4th ed.), Ch. 27, W. H. Freeman; Lilley (1995) *DNA-Protein: Structural Interactions*, IRL Press at Oxford University Press.

[0020] The helix-turn-helix domain consists of two α-helices separated by a short turn. One helix binds to recognition elements within the major groove of DNA, and the other helps to keep the binding helix properly positioned with respect to the rest of the molecule. The helix-turn-helix domain is commonly found in repressor proteins and is typically approximately 20 amino acids long. The helix-turn-helix domain was first identified as a feature of the crystal structure of the bacteriophage λ Cro protein. The structure of this small regulatory protein contained two α-helices separated by 34 Å—the pitch of a DNA double helix. Model building studies showed that these two α-helices would fit into two successive major grooves. In eukaryotes, the helix-turn-helix domain comprises three helices, of which one (the recognition helix) contains the DNA binding region. Proteins having one or more helix-turn-helix domains include, but are not limited to, homeo domain factors (e.g., Antp, Ubx, Engrailed, Eve), POU domain factors (e.g., Oct-1, Oct-2), and developmental regulators (e.g., Forkhead, Myb).

[0021] The leucine zipper or bZIP domain comprises an α-helix that contains a heptad repeat (i.e., at every seventh residue) of leucine residues (or other small, hydrophobic amino acids such as, e.g., isoleucine and/or valine). The leucine zipper is an important feature of many eukaryotic regulatory domains. When a leucine residue occurs every seventh position of an α-helix, the aliphatic side-chains are all oriented on the same side of the helix and they can interact with another helix to form a coiled-coil type of structure. The GCN4 transcription activator in yeast is an example of a leucine zipper motif-containing protein in which the leucine zipper helps to position the two basic regions of the GCN4 dimer to the DNA recognition sequence. Proteins having one or more leucine zipper domains include, but are not limited to, AP-1(-like) components (e.g., Jun, Fos), AP-1(-like) (e.g. GCN4), CRE-BP/ATF, CREB (e.g., CREB, ATF-1), C/EBP-like factors, cell-cycle controlling factors (e.g., Myc, Max), and many viral fusion proteins.

[0022] The helix-loop-helix domain is a variation of the leucine zipper domain. The helix-loop-helix domain is characterized by two α-helices connected by a loop. One helix is typically smaller then the other and, due to the flexibility of the loop, allows dimerization by folding and packing against another helix. The larger of the two helices typically contains the DNA binding region(s). Proteins having one or more

helix-loop-helix domains include, but are not limited to, myogenic transcription factors, and cell-cycle controlling factors (e.g., Myc, Max).

[0023] The winged helix domain typically comprises about 110 amino acids and includes four helices and a two-strand β-sheet. The winged helix turn helix domain is typically 85-90 amino acids long and comprises a three helix bundle and a four-strand β-sheet (wing). Proteins having a winged helix domain include the Forkhead box (FOX) proteins.

[0024] The zinc finger domain is common in eukaryotic DNA-binding proteins, and was first discovered in the eukaryotic transcription factor TFIIIA. The zinc finger domain can coordinate one or more zinc ions to help stabilize its folds. Zinc finger domains can be classified into several different structural families and typically function as interaction modules that bind DNA, RNA, proteins or small molecules. Zinc fingers chelate zinc ions with a combination of cysteine and histidine residues. They can be classified by the type and order of these zinc coordinating residues (e.g. Cys_2His_2, Cys_4, and Cys_6). A more systematic method classifies them into different "fold groups" based on the overall shape of the protein backbone in the folded domain. The most common fold groups of zinc fingers are the Cys_2His_2-like (the "classic" zinc finger), treble clef and zinc ribbon. Zinc finger domains can bind the major groove of DNA.

[0025] The immunoglobulin fold domain comprises a β-sheet structure having large connecting loops which recognize DNA major grooves. Immunoglobulin fold domains are found in immunoglobulin proteins as well as in STAT proteins of the cytokine pathway.

[0026] The B3 domain is approximately 100-120 residues and is found in transcription factors from higher plants. The B3 domain comprises seven β-sheets and two α-helices, which form a pseudo-barrel protein fold. Proteins containing B3 domains are found in higher plants and include auxin response factors (ARFs), abscisic acid insensitive 3 (ABI3) and related to ABI3/VP1 (RAV).

[0027] The TBP-binding domain is found in the TATA-box binding protein, which is a subunit of the eukaryotic transcription factor TFIID. The TBP-binding domain binds the minor groove of DNA.

[0028] As used herein, the term "DNA modifying domain" is intended to refer, but is not limited to, a polypeptide sequence that can modify one or more target nucleosides of a DNA sequence. In certain exemplary embodiments, DNA modifying domains from one or more DNA modifying proteins are provided.

[0029] Proteins having DNA modifying domains are well known in the art and include, but are not limited to, transferases (e.g., terminal deoxynucleotidyl transferase), RNases (RNase A, ribonuclease H), DNases (DNase I), ligases (T4 DNA ligase, *E. coli* DNA ligase), nucleases (51 nuclease), kinases (T4 polynucleotide kinase), phoshatases (calf intestinal alkaline phosphatase, bacterial alkaline phosphatase), exonucleases (X exonuclease), endonucleases, glycosylases (uracil DNA glycosylases), deaminases and the like. A variety of proteins having one or more DNA modifying domains are commercially available (New England Biolabs, Beverly, Mass.; Invitrogen, Carlsbad, Calif.; Sigma-Aldrich, St. Louis, Mo.).

[0030] In certain exemplary embodiments, DNA modifying domains from one or more deaminases are provided. As used herein, the term "deaminase" is intended to include, but is not limited to, a protein that belongs to a class of enzymes that remove one or more amine groups from a target molecule. Deaminases include, but are not limited to, adenosine deaminase, adenine deaminase, cytidine (activation-induced) deaminase, cytosine deaminase, phenylalanine deaminase, uracil deaminase and thymidine deaminase.

[0031] In certain exemplary embodiments, the DNA modifying domain includes activation-induced (cytidine) deaminase (AID) or a portion thereof. AID, a member of the AID/apolipoprotein B RNA Editing Catalytic Component (APOBEC) family, is a 24 kDa enzyme that removes the amino group from the cytidine base in DNA (Delker, R. K., Fugmann, S. D. & Papavasiliou, F. N. A coming-of-age story: activation-induced cytidine deaminase turns 10. *Nat Immunol* 10, 1147-1153 (2009)). It is selectively expressed in the activated B cells in germinal centers (Muramatsu, M., et al. Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells. *J Biol Chem* 274, 18470-18476 (1999)) and is involved in the initiation of three separate immunoglobulin (Ig) diversification processes: somatic hypermutation (SHM), class switch recombination (CSR) and gene-conversion (GC) (Stavnezer, J., Guikema, J. E. & Schrader, C. E. Mechanism and regulation of class switch recombination. *Annu Rev Immunol* 26, 261-292 (2008); Storb, U., et al. Targeting of AID to immunoglobulin genes. *Adv Exp Med Biol* 596, 83-91 (2007); Teng, G. & Papavasiliou, F. N. Immunoglobulin somatic hypermutation. *Annu Rev Genet.* 41, 107-120 (2007)).

[0032] In vitro, AID can deaminate cytidine in ssDNA (Bransteitter, R., Pham, P., Scharff, M. D. & Goodman, M. F. Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. *Proc Natl Acad Sci USA* 100, 4102-4107 (2003)), transcribed dsDNA (Ramiro, A. R., Stavropoulos, P., Jankovic, M. & Nussenzweig, M. C. Transcription enhances AID-mediated cytidine deamination by exposing single-stranded DNA on the nontemplate strand. *Nat Immunol* 4, 452-456 (2003)) and supercoiled dsDNA (Shen, H. M. & Storb, U. Activation-induced cytidine deaminase (AID) can target both DNA strands when the DNA is supercoiled. *Proc Natl Acad Sci USA* 101, 12997-13002 (2004)). In the physiological condition, AID deaminates cytidine, creating uridine:guanosine (U:G) mismatches. The resultant U:G (U=uridine) mismatch is either converted by replication to T:A and C:G base pairs; or the U is removed by an N-glycosylase (UDG) and processed further though Base Excision Repair (BER) pathway; or this mismatch is repaired though Mismatch Repair (MMR) pathway (Peled, J. U., et al. The biochemistry of somatic hypermutation. *Annu Rev Immunol* 26, 481-511 (2008)).

[0033] As used herein, the terms "bind," "binding," "interact," "interacting," "occupy" and "occupying" refer to covalent interactions, noncovalent interactions and steric interactions. A covalent interaction is a chemical linkage between two atoms or radicals formed by the sharing of a pair of electrons (a single bond), two pairs of electrons (a double bond) or three pairs of electrons (a triple bond). Covalent interactions are also known in the art as electron pair interactions or electron pair bonds. Noncovalent interactions include, but are not limited to, van der Waals interactions, hydrogen bonds, weak chemical bonds (via short-range noncovalent forces), hydrophobic interactions, ionic bonds and the like. A review of noncovalent interactions can be found in Alberts et al., in *Molecular Biology of the Cell*, 3d edition, Garland Publishing, 1994. Steric interactions are generally

understood to include those where the structure of the compound is such that it is capable of occupying a site by virtue of its three dimensional structure, as opposed to any attractive forces between the compound and the site.

[0034] As used herein, a "functional fragment" refers to a protein, polypeptide and/or nucleic acid sequence that is not identical to a full-length reference protein, polypeptide or nucleic acid sequence, yet retains the same or similar function as the full-length reference protein, polypeptide or nucleic acid. A functional fragment can possess more, fewer, or the same number of amino acids or nucleic acids as the full-length reference protein, polypeptide or nucleic acid, and/or can contain one or more amino acid or nucleic acid substitutions. Methods for determining the function of a nucleic acid (e.g., coding function, ability to hybridize to another nucleic acid and the like) are well-known in the art (Sambrook et al. Molecular Cloning: A Laboratory Manual, Second edition, Cold Spring Harbor Laboratory Press, 1989 and Third edition, 2001; Ausubel et al., Current Protocols in Molecular Biology, John Wiley & Sons, New York, 1987 and periodic updates; the series Methods in Enzymology, Academic Press, San Diego). Methods for determining protein function are also well-known. Id. For example, the DNA-binding function of a polypeptide can be determined, for example, by filter-binding, electrophoretic mobility-shift, or immunoprecipitation assays. DNA cleavage can be assayed by gel electrophoresis. The ability of a protein to interact with another protein can be determined, for example, by co-immunoprecipitation, chemical cross-linking, two-hybrid assays, complementation (e.g., genetic and/or biochemical) and the like. (See, for example, Fields et al. (1989) *Nature* 340:245-246; U.S. Pat. No. 5,585,245 and PCT WO 98/44350.)

[0035] Methods for designing and constructing fusion proteins (and polynucleotides encoding same) are well known in the art. For example, methods for the design and construction of fusion protein comprising zinc finger proteins (and polynucleotides encoding same) are described in co-owned U.S. Pat. Nos. 6,453,242 and 6,534,261. In certain embodiments, polynucleotides encoding such fusion proteins are constructed. These polynucleotides can be inserted into a vector and the vector can be introduced into a cell as described further herein.

[0036] As used herein, the term "amino acid" includes organic compounds containing both a basic amino group and an acidic carboxyl group. Included within this term are natural amino acids (e.g., L-amino acids), modified and unusual amino acids (e.g., D-amino acids and β-amino acids), as well as amino acids which are known to occur biologically in free or combined form but usually do not occur in proteins. Natural protein occurring amino acids include alanine, arginine, asparagine, aspartic acid, cysteine, glutamic acid, glutamine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, serine, threonine, tyrosine, tryptophan, proline, and valine. Natural non-protein amino acids include arginosuccinic acid, citrulline, cysteine sulfinic acid, 3,4-dihydroxyphenylalanine, homocysteine, homoserine, ornithine, 3-monoiodotyrosine, 3,5-diiodotryosine, 3,5,5,-triiodothyronine, and 3,3',5,5'-tetraiodothyronine. Modified or unusual amino acids include D-amino acids, hydroxylysine, 4-hydroxyproline, N-Cbz-protected amino acids, 2,4-diaminobutyric acid, homoarginine, norleucine, N-methylaminobutyric acid, naphthylalanine, phenylglycine, α-phenylproline, tert-leucine, 4-aminocyclohexylalanine, N-methylnorleucine, 3,4-dehydroproline, N,N-dimethylaminoglycine, N-methylaminoglycine, 4-aminopiperidine-4-carboxylic acid, 6-aminocaproic acid, trans-4-(aminomethyl)-cyclohexanecarboxylic acid, 2-, 3-, and 4-(aminomethyl)-benzoic acid, 1-amino cyclopentane carboxylic acid, 1-aminocyclopropanecarboxylic acid, and 2-benzyl-5-aminopentanoic acid.

[0037] As used herein, the term "peptide" includes compounds that consist of two or more amino acids that are linked by means of a peptide bond. Peptides may have a molecular weight of less than 10,000 Daltons, less than 5,000 Daltons, or less than 2,500 Daltons. The term "peptide" also includes compounds containing both peptide and non-peptide components, such as pseudopeptide or peptidomimetic residues or other non-amino acid components. Such compounds containing both peptide and non-peptide components may also be referred to as a "peptide analog."

[0038] As used herein, the term "protein" includes compounds that consist of amino acids arranged in a linear chain and joined together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues.

[0039] The term "nucleoside," as used herein, includes the natural nucleosides, including 2'-deoxy and 2'-hydroxyl forms, e.g. as described in Komberg and Baker, *DNA Replication,* 2nd Ed. (Freeman, San Francisco, 1992). "Analogs" in reference to nucleosides includes synthetic nucleosides having modified base moieties and/or modified sugar moieties, e.g., described by Scheit, *Nucleotide Analogs* (John Wiley, New York, 1980); Uhlman and Peyman, *Chemical Reviews,* 90:543-584 (1990), or the like, with the proviso that they are capable of specific hybridization. Such analogs include synthetic nucleosides designed to enhance binding properties, reduce complexity, increase specificity, and the like. Polynucleotides comprising analogs with enhanced hybridization or nuclease resistance properties are described in Uhlman and Peyman (cited above); Crooke et al., *Exp. Opin. Ther. Patents,* 6: 855-870 (1996); Mesmaeker et al., *Current Opinion in Structural Biology,* 5:343-355 (1995); and the like. Exemplary types of polynucleotides that are capable of enhancing duplex stability include oligonucleotide phosphoramidates (referred to herein as "amidates"), peptide nucleic acids (referred to herein as "PNAs"), oligo-2'-O-alkylribonucleotides, polynucleotides containing C-5 propynylpyrimidines, locked nucleic acids (LNAs), and like compounds. Such oligonucleotides are either available commercially or may be synthesized using methods described in the literature.

[0040] "Oligonucleotide" or "polynucleotide," which are used synonymously, means a linear polymer of natural or modified nucleosidic monomers linked by phosphodiester bonds or analogs thereof. The term "oligonucleotide" usually refers to a shorter polymer, e.g., comprising from about 3 to about 100 monomers, and the term "polynucleotide" usually refers to longer polymers, e.g., comprising from about 100 monomers to many thousands of monomers, e.g., 10,000 monomers, or more. Oligonucleotides comprising probes or primers usually have lengths in the range of from 12 to 60 nucleotides, and more usually, from 18 to 40 nucleotides. Oligonucleotides and polynucleotides may be natural or synthetic. Oligonucleotides and polynucleotides include deoxyribonucleosides, ribonucleosides, and non-natural analogs thereof, such as anomeric forms thereof, peptide nucleic acids (PNAs), and the like, provided that they are capable of specifically binding to a target genome by way of a regular pattern of monomer-to-monomer interactions, such as Watson-Crick type of base pairing, base stacking, Hoogsteen or reverse Hoogsteen types of base pairing, or the like.

[0041] Usually nucleosidic monomers are linked by phosphodiester bonds. Whenever an oligonucleotide is represented by a sequence of letters, such as "ATGCCTG," it will be understood that the nucleotides are in 5' to 3' order from left to right and that "A" denotes deoxyadenosine, "C" denotes deoxycytidine, "G" denotes deoxyguanosine, "T" denotes deoxythymidine, and "U" denotes the ribonucleoside, uridine, unless otherwise noted. Usually oligonucleotides comprise the four natural deoxynucleotides; however, they may also comprise ribonucleosides or non-natural nucleotide analogs. It is clear to those skilled in the art when oligonucleotides having natural or non-natural nucleotides may be employed in methods and processes described herein. For example, where processing by an enzyme is called for, usually oligonucleotides consisting solely of natural nucleotides are required. Likewise, where an enzyme has specific oligonucleotide or polynucleotide substrate requirements for activity, e.g., single stranded DNA, RNA/DNA duplex, or the like, then selection of appropriate composition for the oligonucleotide or polynucleotide substrates is well within the knowledge of one of ordinary skill, especially with guidance from treatises, such as Sambrook et al., *Molecular Cloning*, Second Edition (Cold Spring Harbor Laboratory, New York, 1989), and like references. Oligonucleotides and polynucleotides may be single stranded or double stranded.

[0042] Oligonucleotides and polynucleotides may optionally include one or more non-standard nucleotide(s), nucleotide analog(s) and/or modified nucleotides. Examples of modified nucleotides include, but are not limited to diaminopurine, S^2T, 5-fluorouracil, 5-bromouracil, 5-chlorouracil, 5-iodouracil, hypoxanthine, xantine, 4-acetylcytosine, 5-(carboxyhydroxylmethyl)uracil, 5-carboxymethylaminomethyl-2-thiouridine, 5-carboxymethylaminomethyluracil, dihydrouracil, beta-D-galactosylqueosine, inosine, N6-isopentenyladenine, 1-methylguanine, 1-methylinosine, 2,2-dimethylguanine, 2-methyladenine, 2-methylguanine, 3-methylcytosine, 5-methylcytosine, N6-adenine, 7-methylguanine, 5-methylaminomethyluracil, 5-methoxyaminomethyl-2-thiouracil, beta-D-mannosylqueosine, 5'-methoxycarboxymethyluracil, 5-methoxyuracil, 2-methylthio-D46-isopentenyladenine, uracil-5-oxyacetic acid (v), wybutoxosine, pseudouracil, queosine, 2-thiocytosine, 5-methyl-2-thiouracil, 2-thiouracil, 4-thiouracil, 5-methyluracil, uracil-5-oxyacetic acid methylester, uracil-5-oxyacetic acid (v), 5-methyl-2-thiouracil, 3-(3-amino-3-N2-carboxypropyl) uracil, (acp3)w, 2,6-diaminopurine and the like. Nucleic acid molecules may also be modified at the base moiety (e.g., at one or more atoms that typically are available to form a hydrogen bond with a complementary nucleotide and/or at one or more atoms that are not typically capable of forming a hydrogen bond with a complementary nucleotide), sugar moiety or phosphate backbone.

[0043] In certain exemplary embodiments, the fusion proteins and methods of targeting DNA modification described herein are used in gene therapy. In certain aspects, stem cell therapy is used to precisely correct inherited point mutations, and then transplant the functionally corrected stem cell back to the patients (Cathomen, T. & Joung, J. K. Zinc-finger nucleases: the next generation emerges. *Mol Ther* 16, 1200-1207 (2008)). Moreover, the fusion proteins and methods described herein can be used as a therapy for cellular proliferative disorders to target oncogenes or non-oncogene addiction (NOA) genes in vivo (Luo, J., Solimini, N. L. & Elledge, S. J. Principles of cancer therapy: oncogene and non-oncogene addiction. *Cell* 136, 823-837 (2009)). High-throughput screens for small molecules that block the activity of oncogenes has been practiced for years, but the art still suffers from a severe lack of clinically effective inhibitors. In certain aspects, the fusion proteins described herein are used to precisely introduce a premature stop codon in the oncogenes (CAG, CAA, CGA to UAG, UAA, UGA, respectively), thus blocking the pathway on which tumor cell depends for its sustained proliferation and survival.

[0044] Cellular proliferative disorders are intended to include disorders associated with rapid proliferation. As used herein, the term "cellular proliferative disorder" includes disorders characterized by undesirable or inappropriate proliferation of one or more subset(s) of cells in a multicellular organism. The term "cancer" refers to various types of malignant neoplasms, most of which can invade surrounding tissues, and may metastasize to different sites (see, for example, PDR Medical Dictionary 1st edition (1995), incorporated herein by reference in its entirety for all purposes). The terms "neoplasm" and "tumor" refer to an abnormal tissue that grows by cellular proliferation more rapidly than normal. Id. Such abnormal tissue shows partial or complete lack of structural organization and functional coordination with the normal tissue which may be either benign (i.e., benign tumor) or malignant (i.e., malignant tumor).

[0045] Examples of the types of neoplasms intended to be encompassed by the present invention include but are not limited to those neoplasms associated with cancers of neural tissue, blood forming tissue, breast, skin, bone, prostate, ovaries, uterus, cervix, liver, lung, brain, larynx, gallbladder, pancreas, rectum, parathyroid, thyroid, adrenal gland, immune system, head and neck, colon, stomach, bronchi, and/or kidneys.

[0046] In certain exemplary embodiments, the fusion proteins and methods of targeting DNA modification described herein are used for constructing transgenic organisms to recapitulate disease. In certain aspects, multiple site modifications are used for the systematic study of common diseases. Particularly, more than 30% of single base changes that have been detected as a cause of genetic disease have occurred at 5'-CpG-3' sites (Holliday, R. & Grigg, G. W. DNA methylation and mutation. Mutat Res 285, 61-67 (1993)). In certain aspects, one or more fusion proteins can be introduced into a cell to make C to T mutations at those sites to generate one or more disease models. In other aspects, single fusion proteins can simultaneously target many repetitive sites in the genome.

[0047] Methods for generating transgenic animals via embryo manipulation and microinjection, particularly animals such as mice, have become conventional in the art and are described, for example, in U.S. Pat. Nos. 4,736,866 and 4,870,009, in U.S. Pat. No. 4,873,191 by Wagner et al., in Hogan, B., Manipulating the Mouse Embryo, (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1986), and in Wilmut et al. (1997) *Nature* 385:810. Similar methods are used for production of other transgenic animals. Methods for producing transgenic non-humans animals that contain selected systems which allow for regulated expression of the transgene are described in Lakso et al. (1992) *Proc. Natl. Acad. Sci. U.S.A.* 89:6232; and O'Gorman et al. (1991) *Science* 251:1351).

[0048] In certain exemplary embodiments, repetitive targets such as endogenous retroviruses are studies using the fusion proteins described herein. In other exemplary embodi-

ments, methods of eliminating mutagenic genomic elements or changing the genetic code genome-wide to make multi-virus resistant cells, which are two examples of needs for potentially thousands of targeted events per genome, using the fusion proteins described herein are provided.

[0049] Viruses include, but are not limited to, DNA or RNA animal viruses. As used herein, RNA viruses include, but are not limited to, virus families such as Picornaviridae (e.g., polioviruses), Reoviridae (e. g., rotaviruses), Togaviridae (e.g., encephalitis viruses, yellow fever virus, rubella virus), Orthomyxoviridae (e.g., influenza viruses), Paramyxoviridae (e.g., respiratory syncytial virus, measles virus, mumps virus, parainfluenza virus), Rhabdoviridae (e.g., rabies virus), Coronaviridae, Bunyaviridae, Flaviviridae, Filoviridae, Arenaviridae, Bunyaviridae and Retroviridae (e.g., human T cell lymphotropic viruses (HTLV), human immunodeficiency viruses (HIV)). As used herein, DNA viruses include, but are not limited to, virus families such as Papovaviridae (e.g., papilloma viruses), Adenoviridae (e.g., adenovirus), Herpesviridae (e.g., herpes simplex viruses), and Poxyiridae (e.g., variola viruses).

[0050] In certain exemplary embodiments, a genome-wide study of the function of retrotransposons in human cells will be performed. Despite their abundance in the human genome (42% of human genome (Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet.* 10, 691-703 (2009)), retrotansopsons have not been thoroughly investigated due to the limitations of current available technologies. By targeting critical and identical elements of retrotransposons, the fusion proteins described herein can inactivate many retrotransposons at the same time, thus revealing their functions.

[0051] In certain exemplary embodiments, vectors such as, for example, expression vectors, containing a nucleic acid encoding one or more fusion proteins described herein are provided. As used herein, the term "vector" refers to a nucleic acid molecule capable of transporting another nucleic acid to which it has been linked. One type of vector is a "plasmid," which refers to a circular double stranded DNA loop into which additional DNA segments can be ligated. Another type of vector is a viral vector, wherein additional DNA segments can be ligated into the viral genome. Certain vectors are capable of autonomous replication in a host cell into which they are introduced (e.g., bacterial vectors having a bacterial origin of replication and episomal mammalian vectors). Other vectors (e.g., non-episomal mammalian vectors) are integrated into the genome of a host cell upon introduction into the host cell, and thereby are replicated along with the host genome. Moreover, certain vectors are capable of directing the expression of genes to which they are operatively linked. Such vectors are referred to herein as "expression vectors." In general, expression vectors of utility in recombinant DNA techniques are often in the form of plasmids. In the present specification, "plasmid" and "vector" can be used interchangeably. However, the invention is intended to include such other forms of expression vectors, such as viral vectors (e.g., replication defective retroviruses, adenoviruses and adeno-associated viruses), which serve equivalent functions.

[0052] In certain exemplary embodiments, the recombinant expression vectors comprise a nucleic acid sequence (e.g., a nucleic acid sequence encoding one or more fusion proteins described herein) in a form suitable for expression of the nucleic acid sequence in a host cell, which means that the recombinant expression vectors include one or more regulatory sequences, selected on the basis of the host cells to be used for expression, which is operatively linked to the nucleic acid sequence to be expressed. Within a recombinant expression vector, "operably linked" is intended to mean that the nucleotide sequence encoding one or more fusion proteins is linked to the regulatory sequence(s) in a manner which allows for expression of the nucleotide sequence (e.g., in an in vitro transcription/translation system or in a host cell when the vector is introduced into the host cell). The term "regulatory sequence" is intended to include promoters, enhancers and other expression control elements (e.g., polyadenylation signals). Such regulatory sequences are described, for example, in Goeddel; *Gene Expression Technology: Methods in Enzymology* 185, Academic Press, San Diego, Calif. (1990). Regulatory sequences include those which direct constitutive expression of a nucleotide sequence in many types of host cells and those which direct expression of the nucleotide sequence only in certain host cells (e.g., tissue-specific regulatory sequences). It will be appreciated by those skilled in the art that the design of the expression vector can depend on such factors as the choice of the host cell to be transformed, the level of expression of protein desired, and the like. The expression vectors described herein can be introduced into host cells to thereby produce proteins or portions thereof, including fusion proteins or portions thereof, encoded by nucleic acids as described herein (e.g., one or more fusion proteins).

[0053] In certain exemplary embodiments, nucleic acid molecules described herein can be inserted into vectors and used as gene therapy vectors. Gene therapy vectors can be delivered to a subject by, for example, intravenous injection, local administration (see, e.g., U.S. Pat. No. 5,328,470), or by stereotactic injection (see, e.g., Chen et al. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91:3054). The pharmaceutical preparation of the gene therapy vector can include the gene therapy vector in an acceptable diluent, or can comprise a slow release matrix in which the gene delivery vehicle is imbedded. Alternatively, where the complete gene delivery vector can be produced intact from recombinant cells, e.g., retroviral vectors, adeno-associated virus vectors, and the like, the pharmaceutical preparation can include one or more cells which produce the gene delivery system (See Gardlik et al. (2005) *Med. Sci. Mon.* 11:110; Salmons and Gunsberg (1993) *Hu. Gene Ther.* 4:129; and Wang et al. (2005) *J. Virol.* 79:10999 for reviews of gene therapy vectors).

[0054] Recombinant expression vectors of the invention can be designed for expression of one or more encoding one or more fusion proteins in prokaryotic or eukaryotic cells. For example, one or more vectors encoding one or more prehairpin intermediate conformations of a fusion protein can be expressed in bacterial cells such as *E. coli*, insect cells (e.g., using baculovirus expression vectors), yeast cells or mammalian cells. Suitable host cells are discussed further in Goeddel, *Gene Expression Technology: Methods in Enzymology* 185, Academic Press, San Diego, Calif. (1990). Alternatively, the recombinant expression vector can be transcribed and translated in vitro, for example using T7 promoter regulatory sequences and T7 polymerase.

[0055] Expression of proteins in prokaryotes is most often carried out in *E. coli* with vectors containing constitutive or inducible promoters directing the expression of either fusion or non-fusion proteins. Fusion vectors add a number of amino acids to a protein encoded therein, usually to the amino ter-

minus of the recombinant protein. Such fusion vectors typically serve three purposes: 1) to increase expression of recombinant protein; 2) to increase the solubility of the recombinant protein; and 3) to aid in the purification of the recombinant protein by acting as a ligand in affinity purification. Often, in fusion expression vectors, a proteolytic cleavage site is introduced at the junction of the fusion moiety and the recombinant protein to enable separation of the recombinant protein from the fusion moiety subsequent to purification of the fusion protein. Such enzymes, and their cognate recognition sequences, include Factor Xa, thrombin and enterokinase. Typical fusion expression vectors include pGEX (Pharmacia Biotech Inc; Smith, D. B. and Johnson, K. S. (1988) *Gene* 67:31-40); pMAL (New England Biolabs, Beverly, Mass.); and pRIT5 (Pharmacia, Piscataway, N.J.) which fuse glutathione S-transferase (GST), maltose E binding protein, or protein A, respectively, to the target recombinant protein.

[0056] In another embodiment, the expression vector encoding one or more fusion proteins is a yeast expression vector. Examples of vectors for expression in yeast *S. cerevisiae* include pYepSec 1 (Baldari, et. al., (1987) *EMBO J.* 6:229-234); pMFa (Kurjan and Herskowitz, (1982) *Cell* 30:933-943); pJRY88 (Schultz et al., (1987) *Gene* 54:113-123); pYES2 (Invitrogen Corporation, San Diego, Calif.); and picZ (Invitrogen Corporation).

[0057] Alternatively, one or more fusion proteins can be expressed in insect cells using baculovirus expression vectors. Baculovirus vectors available for expression of proteins in cultured insect cells (e.g., Sf9 cells) include the pAc series (Smith et al. (1983) *Mol. Cell. Biol.* 3:2156-2165) and the pVL series (Lucklow and Summers (1989) *Virology* 170:31-39).

[0058] In certain exemplary embodiments, one or more fusion proteins herein are expressed in mammalian cells using a mammalian expression vector. Examples of mammalian expression vectors include pCDM8 (Seed, B. (1987) *Nature* 329:840) and pMT2PC (Kaufman et al. (1987) *EMBO J.* 6:187-195). When used in mammalian cells, the expression vector's control functions are often provided by viral regulatory elements. For example, commonly used promoters are derived from polyoma, adenovirus 2, cytomegalovirus and simian virus 40. For other suitable expression systems for both prokaryotic and eukaryotic cells see chapters 16 and 17 of Sambrook, J., Fritsh, E. F., and Maniatis, T. *Molecular Cloning: A Laboratory Manual.* 2nd, ed., Cold Spring Harbor Laboratory, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1989.

[0059] In certain exemplary embodiments, the recombinant mammalian expression vector is capable of directing expression of the nucleic acid preferentially in a particular cell type (e.g., tissue-specific regulatory elements are used to express the nucleic acid). Tissue-specific regulatory elements are known in the art. Non-limiting examples of suitable tissue-specific promoters include lymphoid-specific promoters (Calame and Eaton (1988) *Adv. Immunol.* 43:235), in particular promoters of T cell receptors (Winoto and Baltimore (1989) *EMBO J.* 8:729) and immunoglobulins (Banerji et al. (1983) *Cell* 33:729; Queen and Baltimore (1983) *Cell* 33:741), neuron-specific promoters (e.g., the neurofilament promoter; Byrne and Ruddle (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86:5473), pancreas-specific promoters (Edlund et al. (1985) *Science* 230:912), and mammary gland-specific promoters (e.g., milk whey promoter; U.S. Pat. No. 4,873,316 and European Application Publication No. 264,166). Developmentally-regulated promoters are also encompassed, for example the murine hox promoters (Kessel and Gruss (1990) *Science* 249:374) and the α-fetoprotein promoter (Campes and Tilghman (1989) *Genes Dev.* 3:537).

[0060] In certain exemplary embodiments, host cells into which a recombinant expression vector of the invention has been introduced are provided. The terms "host cell" and "recombinant host cell" are used interchangeably herein. It is understood that such terms refer not only to the particular subject cell but to the progeny or potential progeny of such a cell. Because certain modifications may occur in succeeding generations due to either mutation or environmental influences, such progeny may not, in fact, be identical to the parent cell, but are still included within the scope of the term as used herein.

[0061] A host cell can be any prokaryotic or eukaryotic cell. For example, one or more fusion proteins can be expressed in bacterial cells such as *E. coli*, viral cells such as retroviral cells, insect cells, yeast or mammalian cells (such as Chinese hamster ovary cells (CHO) or COS cells). In other aspects, a host cell is a stem cell. Other suitable host cells are known to those skilled in the art.

[0062] As used herein, the terms "subject," "individual" and "host" are intended to include living organisms such as mammals. Examples of subjects and hosts include, but are not limited to, horses, cows, sheep, pigs, goats, dogs, cats, rabbits, guinea pigs, rats, mice, gerbils, non-human primates (e.g., macaques), humans and the like, non-mammals, including, e.g., non-mammalian vertebrates, such as birds (e.g., chickens or ducks) fish or frogs (e.g., *Xenopus*), and non-mammalian invertebrates, as well as transgenic species thereof.

[0063] As used herein, a "biological sample" may be a single cell or many cells. A biological sample may comprise a single cell type or a combination of two or more cell types. A biological sample further includes a collection of cells that perform a similar function such as those found, for example, in a tissue. Accordingly, certain aspects of the invention are directed to biological samples containing one or more tissues. As used herein, a tissue includes, but is not limited to, epithelial tissue (e.g., skin, the lining of glands, bowel, skin and organs such as the liver, lung, kidney), endothelium (e.g., the lining of blood and lymphatic vessels), mesothelium (e.g., the lining of pleural, peritoneal and pericardial spaces), mesenchyme (e.g., cells filling the spaces between the organs, including fat, muscle, bone, cartilage and tendon cells), blood cells (e.g., red and white blood cells), neurons, germ cells (e.g., spermatozoa, oocytes), placenta, stem cells and the like. A tissue sample includes microscopic samples as well as macroscopic samples.

[0064] Delivery of nucleic acids described herein (e.g., vector DNA) can be by any suitable method in the art. For example, delivery may be by injection, gene gun, by application of the nucleic acid in a gel, oil, or cream, by electroporation, using lipid-based transfection reagents, or by any other suitable transfection method.

[0065] As used herein, the terms "transformation" and "transfection" are intended to refer to a variety of art-recognized techniques for introducing foreign nucleic acid (e.g., DNA) into a host cell, including calcium phosphate or calcium chloride co-precipitation, DEAE-dextran-mediated transfection, lipofection (e.g., using commercially available reagents such as, for example, LIPOFECTIN® (Invitrogen

Corp., San Diego, Calif.), LIPOFECTAMINE® (Invitrogen), FUGENE® (Roche Applied Science, Basel, Switzerland), JETPEI™ (Polyplus-transfection Inc., New York, N.Y.), EFFECTENE® (Qiagen, Valencia, Calif.), DREAMFECT™ (OZ Biosciences, France) and the like), or electroporation (e.g., in vivo electroporation). Suitable methods for transforming or transfecting host cells can be found in Sambrook, et al. (*Molecular Cloning: A Laboratory Manual.* 2nd, ed., Cold Spring harbor Laboratory, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1989), and other laboratory manuals.

[0066] In certain exemplary embodiments, one or more agents (e.g., one or more fusion proteins or one or more vectors encoding one or more fusion proteins) are provided in a pharmaceutically acceptable carrier. As used herein, the language "pharmaceutically acceptable carrier" is intended to include any and all solvents, dispersion media, coatings, antibacterial and antifungal agents, isotonic and absorption delaying agents, and the like, compatible with pharmaceutical administration. The use of such media and agents for pharmaceutically active substances is well known in the art. Except insofar as any conventional media or agent is incompatible with the active compound, use thereof in the compositions is contemplated. Supplementary active compounds can also be incorporated into the compositions. Pharmaceutically acceptable carriers and their formulations are known to those skilled in the art and described, for example, in Remington's Pharmaceutical Sciences, (19th edition), ed. A. Gennaro, 1995, Mack Publishing Company, Easton, Pa.

[0067] In certain exemplary embodiments, pharmaceutical formulations of a therapeutically effective amount of one or more agents (e.g., one or more fusion proteins or one or more vectors encoding one or more fusion proteins) are administered by intravenous injection, intraperitoneal injection, oral administration or by other parenteral routes (e.g. intradermal, subcutaneous, oral (e.g., inhalation), transdermal (topical), transmucosal, and rectal administration), or by intrathecal and intraventricular injections into the CNS, in an admixture with a pharmaceutically acceptable carrier adapted for the route of administration.

[0068] Solutions or suspensions used for parenteral, intradermal, subcutaneous or central nervous system application can include the following components: a sterile diluent such as water for injection, saline solution, fixed oils, polyethylene glycols, glycerin, propylene glycol or other synthetic solvents; antibacterial agents such as benzyl alcohol or methyl parabens; antioxidants such as ascorbic acid or sodium bisulfite; chelating agents such as ethylenediaminetetraacetic acid; buffers such as acetates, citrates or phosphates and agents for the adjustment of tonicity such as sodium chloride or dextrose. pH can be adjusted with acids or bases, such as hydrochloric acid or sodium hydroxide. The parenteral preparation can be enclosed in ampoules, disposable syringes or multiple dose vials made of glass or plastic.

[0069] Compositions intended for oral use may be prepared in solid or liquid forms according to any method known to the art for the manufacture of pharmaceutical compositions. The compositions may optionally contain sweetening, flavoring, coloring, perfuming, and/or preserving agents in order to provide a more palatable preparation. Solid dosage forms for oral administration include capsules, tablets, pills, powders, and granules. In such solid forms, the active compound is admixed with at least one inert pharmaceutically acceptable carrier or excipient. These may include, for example, inert diluents, such as calcium carbonate, sodium carbonate, lactose, sucrose, starch, calcium phosphate, sodium phosphate, or kaolin. Binding agents, buffering agents, and/or lubricating agents (e.g., magnesium stearate) may also be used. Tablets and pills can additionally be prepared with enteric coatings.

[0070] Pharmaceutical compositions suitable for injectable use include sterile aqueous solutions (where water soluble) or dispersions and sterile powders for the extemporaneous preparation of sterile injectable solutions or dispersion. For intravenous administration, suitable carriers include physiological saline, bacteriostatic water, CREMOPHOR EL™ (BASF, Parsippany, N.J.) or phosphate buffered saline (PBS). In all cases, the composition must be sterile and should be fluid to the extent that easy syringability exists. It must be stable under the conditions of manufacture and storage and must be preserved against the contaminating action of microorganisms such as bacteria and fungi. The carrier can be a solvent or dispersion medium containing, for example, water, ethanol, polyol (for example, glycerol, propylene glycol, and liquid polyethylene glycol, and the like), and suitable mixtures thereof. The proper fluidity can be maintained, for example, by the use of a coating such as lecithin, by the maintenance of the required particle size in the case of dispersion and by the use of surfactants. Prevention of the action of microorganisms can be achieved by various antibacterial and antifungal agents, for example, parabens, chlorobutanol, phenol, ascorbic acid, thimerosal, and the like. In certain exemplary embodiments, isotonic agents, for example, sugars, polyalcohols such as mannitol, sorbitol, and/or sodium chloride, will be included in the composition. Prolonged absorption of the injectable compositions can be brought about by including in the composition an agent which delays absorption, for example, aluminum monostearate and gelatin.

[0071] Sterile injectable solutions can be prepared by incorporating one or more agents (e.g., one or more fusion proteins or one or more vectors encoding one or more fusion proteins) in the required amount in an appropriate solvent with one or a combination of ingredients enumerated above, as required, followed by filtered sterilization. Generally, dispersions are prepared by incorporating the active compound into a sterile vehicle which contains a basic dispersion medium and the required other ingredients from those enumerated above. In the case of sterile powders for the preparation of sterile injectable solutions, exemplary methods of preparation are vacuum drying and freeze-drying which yields a powder of the active ingredient plus any additional desired ingredient from a previously sterile-filtered solution thereof.

[0072] Oral compositions generally include an inert diluent or an edible carrier. They can be enclosed in gelatin capsules or compressed into tablets. For the purpose of oral therapeutic administration, the active compound can be incorporated with excipients and used in the form of tablets, troches, or capsules. Oral compositions can also be prepared using a fluid carrier for use as a mouthwash, wherein the compound in the fluid carrier is applied orally and swished and expectorated or swallowed. Pharmaceutically compatible binding agents, and/or adjuvant materials can be included as part of the composition. The tablets, pills, capsules, troches and the like can contain any of the following ingredients, or compounds of a similar nature: A binder such as microcrystalline cellulose, gum tragacanth or gelatin; an excipient such as starch or lactose, a disintegrating agent such as alginic, acid, Primogel,

or corn starch; a lubricant such as magnesium stearate or Sterotes; a glidant: such as colloidal silicon dioxide; a sweetening agent such as sucrose or saccharin; or a flavoring agent such as peppermint, methyl salicylate, or orange flavoring.

[0073] In one embodiment, one or more agents (e.g., one or more fusion proteins or one or more vectors encoding one or more fusion proteins) are prepared with carriers that will protect the compound against rapid elimination from the body, such as a controlled release formulation, including implants and microencapsulated delivery systems. Biodegradable, biocompatible polymers can be used, such as ethylene vinyl acetate, polyanhydrides, polyglycolic acid, collagen, polyorthoesters, and polylactic acid. Methods for preparation of such formulations will be apparent to those skilled in the art. The materials can also be obtained commercially from Alza Corporation and Nova Pharmaceuticals, Inc. Liposomal suspensions (including liposomes targeted to infected cells with monoclonal antibodies to viral antigens) can also be used as pharmaceutically acceptable carriers. These may be prepared according to methods known to those skilled in the art, for example, as described in U.S. Pat. No. 4,522,811.

[0074] Nasal compositions generally include nasal sprays and inhalants. Nasal sprays and inhalants can contain one or more active components and excipients such as preservatives, viscosity modifiers, emulsifiers, buffering agents and the like. Nasal sprays may be applied to the nasal cavity for local and/or systemic use. Nasal sprays may be dispensed by a non-pressurized dispenser suitable for delivery of a metered dose of the active component. Nasal inhalants are intended for delivery to the lungs by oral inhalation for local and/or systemic use. Nasal inhalants may be dispensed by a closed container system for delivery of a metered dose of one or more active components.

[0075] In one embodiment, nasal inhalants are used with an aerosol. This is accomplished by preparing an aqueous aerosol, liposomal preparation or solid particles containing the compound. A non-aqueous (e.g., fluorocarbon propellant) suspension could be used. Sonic nebulizers may be used to minimize exposing the agent to shear, which can result in degradation of the compound.

[0076] Ordinarily, an aqueous aerosol is made by formulating an aqueous solution or suspension of the agent together with conventional pharmaceutically acceptable carriers and stabilizers. The carriers and stabilizers vary with the requirements of the particular compound, but typically include nonionic surfactants (Tweens, Pluronics, or polyethylene glycol), innocuous proteins like serum albumin, sorbitan esters, oleic acid, lecithin, amino acids such as glycine, buffers, salts, sugars or sugar alcohols. Aerosols generally are prepared from isotonic solutions.

[0077] Systemic administration can also be by transmucosal or transdermal means. For transmucosal or transdermal administration, penetrants appropriate to the barrier to be permeated are used in the formulation. Such penetrants are generally known in the art, and include, for example, for transmucosal administration, detergents, bile salts, and fusidic acid derivatives. Transmucosal administration can be accomplished through the use of nasal sprays or suppositories. For transdermal administration, the active compounds are formulated into ointments, salves, gels, or creams as generally known in the art.

[0078] One or more agents (e.g., one or more fusion proteins or one or more vectors encoding one or more fusion proteins) can also be prepared in the form of suppositories (e.g., with conventional suppository bases such as cocoa butter and other glycerides) or retention enemas for rectal delivery.

[0079] In one embodiment, one or more agents (e.g., one or more fusion proteins or one or more vectors encoding one or more fusion proteins) are prepared with carriers that will protect them against rapid elimination from the body, such as a controlled release formulation, including implants and microencapsulated delivery systems. Biodegradable, biocompatible polymers can be used, such as ethylene vinyl acetate, polyanhydrides, polyglycolic acid, collagen, polyorthoesters, and polylactic acid. Methods for preparation of such formulations will be apparent to those skilled in the art. The materials can also be obtained commercially from Alza Corporation and Nova Pharmaceuticals, Inc. Liposomal suspensions (including liposomes targeted to infected cells with monoclonal antibodies to viral antigens) can also be used as pharmaceutically acceptable carriers. These can be prepared according to methods known to those skilled in the art, for example, as described in U.S. Pat. No. 4,522,811.

[0080] It is especially advantageous to formulate oral, parenteral or CNS direct delivery compositions in dosage unit form for ease of administration and uniformity of dosage. Dosage unit form as used herein refers to physically discrete units suited as unitary dosages for the subject to be treated; each unit containing a predetermined quantity of active compound calculated to produce the desired therapeutic effect in association with the required pharmaceutical carrier. The specification for the dosage unit forms of the invention are dictated by and directly dependent on the unique characteristics of the active compound and the particular therapeutic effect to be achieved, and the limitations inherent in the art of compounding such an active compound for the treatment of individuals.

[0081] Toxicity and therapeutic efficacy of one or more agents (e.g., one or more fusion proteins or one or more vectors encoding one or more fusion proteins) can be determined by standard pharmaceutical procedures in cell cultures or experimental animals, e.g., for determining the LD50 (the dose lethal to 50% of the population) and the ED50 (the dose therapeutically effective in 50% of the population). The dose ratio between toxic and therapeutic effects is the therapeutic index and it can be expressed as the ratio LD50/ED50. Compounds which exhibit large therapeutic indices are preferred. While compounds that exhibit toxic side effects may be used, care should be taken to design a delivery system that targets such compounds to the site of affected tissue in order to minimize potential damage to uninfected cells and, thereby, reduce side effects.

[0082] Data obtained from cell culture assays and/or animal studies can be used in formulating a range of dosage for use in humans. The dosage typically will lie within a range of circulating concentrations that include the ED50 with little or no toxicity. The dosage may vary within this range depending upon the dosage form employed and the route of administration utilized. For any compound used in the method of the invention, the therapeutically effective dose can be estimated initially from cell culture assays. A dose may be formulated in animal models to achieve a circulating plasma concentration range that includes the IC50 (i.e., the concentration of the test compound which achieves a half-maximal inhibition of

symptoms) as determined in cell culture. Such information can be used to more accurately determine useful doses in humans. Levels in plasma may be measured, for example, by high performance liquid chromatography.

[0083] In certain exemplary embodiments, a method for treatment of a disease or disorder described herein includes the step of administering a therapeutically effective amount of an agent (e.e. g., one or more fusion proteins or one or more vectors encoding one or more fusion proteins) to a subject. As defined herein, a therapeutically effective amount of agent (i.e., an effective dosage) ranges from about 0.0001 to 30 mg/kg body weight, from about 0.001 to 25 mg/kg body weight, from about 0.01 to 20 mg/kg body weight, from about 0.1 to 15 mg/kg body weight, or from about 1 to 10 mg/kg, 2 to 9 mg/kg, 3 to 8 mg/kg, 4 to 7 mg/kg, or 5 to 6 mg/kg body weight. The skilled artisan will appreciate that certain factors may influence the dosage required to effectively treat a subject, including but not limited to the severity of the disease or disorder, previous treatments, the general health and/or age of the subject, and other diseases present. Moreover, treatment of a subject with a therapeutically effective amount of one or more agents (e.g., one or more fusion proteins or one or more vectors encoding one or more fusion proteins) can include a single treatment or, in certain exemplary embodiments, can include a series of treatments. It will also be appreciated that the effective dosage of agent used for treatment may increase or decrease over the course of a particular treatment. Changes in dosage may result from the results of diagnostic assays as described herein. The pharmaceutical compositions can be included in a container, pack, or dispenser together with instructions for administration.

[0084] Embodiments of the invention are directed to a first nucleic acid (e.g., a nucleic acid sequence encoding a fusion protein comprising one or more DNA binding domains and/or one or more DNA modifying domains) or polypeptide sequence (e.g., a fusion protein comprising one or more DNA binding domains and/or one or more DNA modifying domains) having a certain sequence identity or percent homology to a second nucleic acid or polypeptide sequence, respectively.

[0085] Techniques for determining nucleic acid and amino acid "sequence identity" are known in the art. Typically, such techniques include determining the nucleotide sequence of genomic DNA, mRNA or cDNA made from an mRNA for a gene and/or determining the amino acid sequence that it encodes, and comparing one or both of these sequences to a second nucleotide or amino acid sequence, as appropriate. In general, "identity" refers to an exact nucleotide-to-nucleotide or amino acid-to-amino acid correspondence of two polynucleotides or polypeptide sequences, respectively. Two or more sequences (polynucleotide or amino acid) can be compared by determining their "percent identity." The percent identity of two sequences, whether nucleic acid or amino acid sequences, is the number of exact matches between two aligned sequences divided by the length of the shorter sequences and multiplied by 100.

[0086] An approximate alignment for nucleic acid sequences is provided by the local homology algorithm of Smith and Waterman, *Advances in Applied Mathematics* 2:482-489 (1981). This algorithm can be applied to amino acid sequences by using the scoring matrix developed by Dayhoff, *Atlas of Protein Sequences and Structure*, M. O. Dayhoff ed., 5 suppl. 3:353-358, National Biomedical Research Foundation, Washington, D.C., USA, and normalized by Gribskov (1986) *Nucl. Acids Res.* 14:6745. An exemplary implementation of this algorithm to determine percent identity of a sequence is provided by the Genetics Computer Group (Madison, Wis.) in the "BestFit" utility application. The default parameters for this method are described in the *Wisconsin Sequence Analysis Package Program Manual*, Version 8 (1995) (available from Genetics Computer Group, Madison, Wis.).

[0087] One method of establishing percent identity in the context of the present invention is to use the MPSRCH package of programs copyrighted by the University of Edinburgh, developed by John F. Collins and Shane S. Sturrok, and distributed by IntelliGenetics, Inc. (Mountain View, Calif.). From this suite of packages, the Smith-Waterman algorithm can be employed where default parameters are used for the scoring table (for example, gap open penalty of 12, gap extension penalty of one, and a gap of six). From the data generated the "match" value reflects "sequence identity." Other suitable programs for calculating the percent identity or similarity between sequences are generally known in the art, for example, another alignment program is BLAST, used with default parameters. For example, BLASTN and BLASTP can be used using the following default parameters: genetic code=standard; filter=none; strand=both; cutoff=60; expect=10; Matrix=BLOSUM62; Descriptions=50 sequences; sort by =HIGH SCORE; Databases=non-redundant, GenBank+EMBL+DDBJ+PDB+GenBank CDS translations+Swiss protein+Spupdate+PIR. Details of these programs can be found at the NCBI/NLM web site.

[0088] Alternatively, homology can be determined by hybridization of polynucleotides under conditions that form stable duplexes between homologous regions, followed by digestion with single-stranded-specific nuclease(s), and size determination of the digested fragments. Two DNA sequences, or two polypeptide sequences are "substantially homologous" to each other when the sequences exhibit at least about 80%-85%, at least about 85%-90%, at least about 90%-95%, or at least about 95%-98% sequence identity over a defined length of the molecules, as determined using the methods above. As used herein, substantially homologous also refers to sequences showing complete identity to the specified DNA or polypeptide sequence. DNA sequences that are substantially homologous can be identified in a Southern hybridization experiment under, for example, stringent conditions, as defined for that particular system. Defining appropriate hybridization conditions is within the skill of the art. See, e.g., Sambrook et al., *Molecular Cloning: A Laboratory Manual*, Second Edition, (1989) Cold Spring Harbor, N.Y.; *Nucleic Acid Hybridization: A Practical Approach*, editors B. D. Hames and S. J. Higgins, (1985) Oxford; Washington, D.C.; IRL Press.

[0089] Two nucleic acid fragments are considered to "selectively hybridize" as described herein. The degree of sequence identity between two nucleic acid molecules affects the efficiency and strength of hybridization events between such molecules. A partially identical nucleic acid sequence will at least partially inhibit a completely identical sequence from hybridizing to a target molecule. Inhibition of hybridization of the completely identical sequence can be assessed using hybridization assays that are well known in the art (e.g., Southern blot, Northern blot, solution hybridization, or the like, see Sambrook, et al., supra). Such assays can be conducted using varying degrees of selectivity, for example, using conditions varying from low to high stringency. If con-

ditions of low stringency are employed, the absence of non-specific binding can be assessed using a secondary probe that lacks even a partial degree of sequence identity (for example, a probe having less than about 30% sequence identity with the target molecule), such that, in the absence of non-specific binding events, the secondary probe will not hybridize to the target.

[0090] When utilizing a hybridization-based detection system, a nucleic acid probe is chosen that is complementary to a target nucleic acid sequence, and then by selection of appropriate conditions the probe and the target sequence "selectively hybridize," or bind, to each other to form a hybrid molecule. A nucleic acid molecule that is capable of hybridizing selectively to a target sequence under "moderately stringent" conditions typically hybridizes under conditions that allow detection of a target nucleic acid sequence of at least about 10-14 nucleotides in length having at least approximately 70% sequence identity with the sequence of the selected nucleic acid probe. Stringent hybridization conditions typically allow detection of target nucleic acid sequences of at least about 10-14 nucleotides in length having a sequence identity of greater than about 90-95% with the sequence of the selected nucleic acid probe. Hybridization conditions useful for probe/target hybridization where the probe and target have a specific degree of sequence identity, can be determined as is known in the art (see, for example, *Nucleic Acid Hybridization*, supra).

[0091] With respect to stringency conditions for hybridization, it is well known in the art that numerous equivalent conditions can be employed to establish a particular stringency by varying, for example, the following factors: the length and nature of probe and target sequences, base composition of the various sequences, concentrations of salts and other hybridization solution components, the presence or absence of blocking agents in the hybridization solutions (e.g., formamide, dextran sulfate, and polyethylene glycol), hybridization reaction temperature and time parameters, as well as varying wash conditions. The selection of a particular set of hybridization conditions is selected following standard methods in the art (see, for example, Sambrook et al., supra).

[0092] As used herein, the term "hybridizes under stringent conditions" is intended to describe conditions for hybridization and washing under which nucleotide sequences at least 60% identical to each other typically remain hybridized to each other. In one aspect, the conditions are such that sequences at least about 70%, at least about 80%, at least about 85% or 90% or more identical to each other typically remain hybridized to each other. Such stringent conditions are known to those skilled in the art and can be found in *Current Protocols in Molecular Biology*, John Wiley & Sons, NY (1989), 6.3.1-6.3.6. A non-limiting example of stringent hybridization conditions are hybridization in 6× sodium chloride/sodium citrate (SSC) at about 45° C., followed by one or more washes in 0.2×SSC, 0.1% SDS at 50° C., at 55° C., or at 60° C. or 65° C.

[0093] A first polynucleotide is "derived from" a second polynucleotide if it has the same or substantially the same base-pair sequence as a region of the second polynucleotide, its cDNA, complements thereof, or if it displays sequence identity as described above. A first polypeptide is derived from a second polypeptide if it is encoded by a first polynucleotide derived from a second polynucleotide, or displays sequence identity to the second polypeptides as described above. In the present invention, when a DNA binding domain and/or a DNA modifying domain is "derived from" a reference protein or polypeptide, the reference protein or polypeptide need not be explicitly produced, it is simply considered to be the original source of the DNA binding domain and/or a DNA modifying domain and/or nucleic acid sequences that encode it. DNA binding domains and/or a DNA modifying domains can, for example, be produced recombinantly or synthetically, by methods known in the art, or alternatively, purified from cell culture.

[0094] In certain aspects, nucleic acid sequences derived or obtained from one or more organisms are provided. As used herein, the term "organism" includes, but is not limited to, a human, a non-human primate, a cow, a horse, a sheep, a goat, a pig, a dog, a cat, a rabbit, a mouse, a rat, a gerbil, a frog, a toad, a fish (e.g., *Danio rerio*) a roundworm (e.g., *C. elegans*) and any transgenic species thereof. The term "organism" further includes, but is not limited to, a yeast (e.g., *S. cerevisiae*) cell, a yeast tetrad, a yeast colony, a bacterium, a bacterial colony, a virion, virosome, virus-like particle and/or cultures thereof, and the like.

[0095] Oligonucleotides or fragments thereof may be purchased from commercial sources. Oligonucleotide sequences may be prepared by any suitable method, e.g., the phosphoramidite method described by Beaucage and Carruthers ((1981) *Tetrahedron Lett.* 22: 1859) or the triester method according to Matteucci et al. (1981) *J. Am. Chem. Soc.* 103:3185), both incorporated herein by reference in their entirety for all purposes, or by other chemical methods using either a commercial automated oligonucleotide synthesizer or high-throughput, high-density array methods described herein and known in the art (see U.S. Pat. Nos. 5,602,244, 5,574,146, 5,554,744, 5,428,148, 5,264,566, 5,141,813, 5,959,463, 4,861,571 and 4,659,774, incorporated herein by reference in its entirety for all purposes). Pre-synthesized oligonucleotides and chips containing oligonucleotides may also be obtained commercially from a variety of vendors.

[0096] In an exemplary embodiment, construction and/or selection oligonucleotides may be synthesized on a solid support using maskless array synthesizer (MAS). Maskless array synthesizers are described, for example, in PCT application No. WO 99/42813 and in corresponding U.S. Pat. No. 6,375,903. Other examples are known of maskless instruments which can fabricate a custom DNA microarray in which each of the features in the array has a single stranded DNA molecule of desired sequence. An exemplary type of instrument is the type shown in FIG. 5 of U.S. Pat. No. 6,375,903, based on the use of reflective optics. It is a desirable that this type of maskless array synthesizer is under software control. Since the entire process of microarray synthesis can be accomplished in only a few hours, and since suitable software permits the desired DNA sequences to be altered at will, this class of device makes it possible to fabricate microarrays including DNA segments of different sequence every day or even multiple times per day on one instrument. The differences in DNA sequence of the DNA segments in the microarray can also be slight or dramatic, it makes no difference to the process. The MAS instrument may be used in the form it would normally be used to make microarrays for hybridization experiments, but it may also be adapted to have features specifically adapted for the compositions, methods, and systems described herein. For example, it may be desirable to substitute a coherent light source, i.e., a laser, for the light source shown in FIG. 5 of the above-mentioned U.S. Pat. No. 6,375,903. If a laser is used as the

light source, a beam expanded and scatter plate may be used after the laser to transform the narrow light beam from the laser into a broader light source to illuminate the micromirror arrays used in the maskless array synthesizer. It is also envisioned that changes may be made to the flow cell in which the microarray is synthesized. In particular, it is envisioned that the flow cell can be compartmentalized, with linear rows of array elements being in fluid communication with each other by a common fluid channel, but each channel being separated from adjacent channels associated with neighboring rows of array elements. During microarray synthesis, the channels all receive the same fluids at the same time. After the DNA segments are separated from the substrate, the channels serve to permit the DNA segments from the row of array elements to congregate with each other and begin to self-assemble by hybridization. Other methods for synthesizing oligonucleotides (e.g., Oligopaints) include, for example, light-directed methods utilizing masks, flow channel methods, spotting methods, pin-based methods, and methods utilizing multiple supports.

[0097] It is to be understood that the embodiments of the present invention which have been described are merely illustrative of some of the applications of the principles of the present invention. Numerous modifications may be made by those skilled in the art based upon the teachings presented herein without departing from the true spirit and scope of the invention. The contents of all references, patents and published patent applications cited throughout this application are hereby incorporated by reference in their entirety for all purposes.

[0098] The following examples are set forth as being representative of the present invention. These examples are not to be construed as limiting the scope of the invention as these and other equivalent embodiments will be apparent in view of the present disclosure, figures, tables, and accompanying claims.

Example I

Sequence Specific DNA Deamination

[0099] In one aspect, a fusion protein can include a DNA binding domain including any of the pairs of zinc fingers targeting EGFP described in Maeder et al. ((2008) *Mol. Cell,* 31:294) or portions thereof functionally linked to a DNA modifying domain including AID or portions thereof AID is a 24 kDa enzyme that removes the amino group from the cytidine base in DNA especially within hotspot motifs (WRCY motifs W=adenosine or thymidine, R=purine, C=cytidine, Y=pyrimidine). It is involved in the initiation of three separate immunoglobulin (Ig) diversification processes, somatic hypermutation (SHM), class switch recombination (CSR) and gene-conversion (GC). AID has been shown in vitro to be active on single stranded DNA, and has been shown to require active transcription in order to exert its deaminating activity. The resultant U:G (U=uridine) mismatch is then either: 1) converted by replication to T:A and C:G base pairs; 2) The U removed by an N-glycosylase and replaced by A,C,G, or T; or 3) error-prone mismatch repair (MMR) in the region. The intrinsic specificity of AID can either be exploited if an appropriate matching site for targeting can be found, or the specificity can be reduced or shifted to another sequence using design principles and the 3D structure of the deaminases.

(SEQ ID NO: 1)
PGERPFQCRICMRNFSXXXXXXXHTRTHTGEKPFQCRICMRNFSXXXXX

XXHLRTHTGEKPFQCRICMRNFSXXXXXXXHLKTH

[0100] The X's represent the recognition helix residues that are given in the Maeder et al. *Mol. Cell* Supplemental table (*Molecular Cell*, Volume 31, Issue 2, 294-301, 25 July 2008, doi:10.1016/j.molcel.2008.06.016).

Example II

Activation-Induced Deaminase and Zinc Finger Protein Fusion Proteins

[0101] The ability to modify a large number of sites in the human genome is very helpful for testing hypotheses derived from genomic sequence data. Current modification methodologies including homologous recombination and zinc finger nuclease-associated homologous recombination are low throughput and are relatively inefficient. The fusion proteins described herein will generate a new gene targeting method. In certain aspects, a fusion protein is provided wherein the DNA modifying domain includes a functional fragment of AID and the DNA binding domain includes a functional fragment of a ZFP. AID is a DNA deaminase that deaminates cytidine to uridine, thus introducing a single nucleotide transition. Customized ZFP can specifically bind to defined sequences. Whether a fusion AID-ZFP retains the activities of its modules and whether this function can be used as a targeting modification tool in the human genome will be ascertained. This question will be examined by (1) testing whether AID-ZFP can deaminate specific cytidine in *Escherichia coli*; (2) assessing the toxicity and off-target rate of AID-ZFP; and (3) testing whether AID-ZFP can introduce specific mutations in the human genome. This method is promising for gene therapy and genome-wide gene engineering.

[0102] The need to modify multiple sites in the genome is rapidly increasing due to the growth of hypotheses flowing from genomic sequence data. Spontaneous homologous recombination is impractical, however, because of its low efficiency (Zeng, X. & Rao, M. S. Controlled genetic modification of stem cells for developing drug discovery tools and novel therapeutic applications. *Curr Opin Mol Ther* 10, 207-213 (2008)). Several new methods have been developed which allow higher efficiency: 1) Introducing single-stranded oligomers in strains with mismatch repair deficiency and over-expression of homologous DNA repairing proteins (Wang, H. H., et al. Programming cells by multiplex genome engineering and accelerated evolution. *Nature* 460, 894-898 (2009)). 2) Using nucleases and recombinases to stimulate homologous recombination (e.g. Zinc Finger Nuclease (ZFN) (Foley, J. E., et al. Rapid mutation of endogenous zebrafish genes using zinc finger nucleases made by Oligomerized Pool ENgineering (OPEN). *PLoS One* 4, e4348 (2009)), meganucleases (Fajardo-Sanchez, E., Stricher, F., Paques, F., Isalan, M. & Serrano, L. Computer design of obligate heterodimer meganucleases allows efficient cutting of custom DNA sequences. *Nucleic Acids Res* 36, 2163-2173 (2008)), phage integrases (Groth, A. C. & Calos, M. P. Phage integrases: biology and applications. *J Mol Biol* 335, 667-678 (2004)) and other microbial recombinases (Id.).

[0103] Importantly, these technologies are limited by several factors. First, as three different molecules (DNA donor, acceptor and protein catalyst) need to be present simultaneously for successful recombination, this requirement limits the efficiency of targeting while also increasing the possibility of random alterations. Second, most of the strategies can cause detrimental DNA lesions. For example, ZFN facilitates gene targeting by introducing double stranded breaks (DSB), which would be repaired by homologous recombination. However, the efficiency of desired low-error replacement of targeted DNA by homologous recombination (HR) is low compared to error-prone non-homologous end joining (NHEJ) and random integration (Kandavelou, K., et al. Targeted manipulation of mammalian genomes using designed zinc finger nucleases. *Biochem Biophys Res Commun* 388, 56-61 (2009)). Estimates of native HR:NHEJ efficiencies vary from 1:30 to 1:40000 (Yanez, R. J. & Porter, A. C. Therapeutic gene targeting. *Gene Ther* 5, 149-159 (1998)). Moreover, the ZFN method is impractical for modifying multiple sites at the same time because different ZFNs would cut the genome to pieces, which would result in one or more chromosome deletion(s), translocation(s), inversion(s) and/or other detrimental effects.

[0104] AID-ZFPs hold great promise as a new tool for targeted mutation. First, AID-ZFPs alone can introduce precise mutations in the genome without the presence of any DNA donor. Second, engineered AID-ZFP would deaminate cytidine without introducing truncations into the genome, making multiple sites modification feasible. Third, the ability to introduce single mutations in the genome makes AID-ZFP useful in many contexts. By changing C to T (or G to A), AID can introduce premature stop codon(s) (CGA, CAA, CAG to TGA, TAA, TAG, respectively), abolish start codon(s) (ATG to ATA); introduce alternative splicing sites (GT - - - AG to (A)T - - - A(A)), change SNP residues, and/or change RNA editing sites (Li, J. B., et al. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* 324, 1210-1213 (2009)).

Example III

AID-ZFP Deamination of Specific Cytidine Residues in the *Escherichia coli* Genome

[0105] A green fluorescent protein (GFP) reporter system incorporated into the genome will be constructed as depicted in FIG. **1**. A group of synthesized, double stranded DNA (dsDNA) fragments will be generated (red). One sequence will have OCT4 ZFP (Hockemeyer, D., et al. Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nat Biotechnol* 27, 851-857 (2009)) binding sites (GAGCAGGCAGGGTCAGCT) (SEQ ID NO:2) in the downstream of a broken start codon "ACG." Another sequence will have a "broken" start codon "ACG" followed by random sequence. Both of these sequences will have a pBAD promoter region and ribosome binding sites at the 5' end and a flexible linker coding region at the 3' end. These pieces of DNA will be constructed between an antibiotic resistant gene (yellow) and a promoter-less green florescent protein gene (gfp) (green) which is in the right translation frame. The final homologous recombination construct will be generated by tagging 50 base pair homologies (A & B) at both ends, and transformed into recombination-proficient cells (Wang, H. H., et al. Programming cells by multiplex genome engineering and accelerated evolution. *Nature* 460, 894-898 (2009)). Recombinants will be selected with the antibiotic resistant marker and be further verified by PCR. As a positive control, a parallel construct with a normal start codon will be incorporated into the genome. Florescent microscopy will be used to examine the expression of GFP.

[0106] If GFP can be expressed in the positive control but not the experiment group, this will indicate that the tagged-GFP can be expressed and is functional. If the GFP fluorescence cannot be detected in the positive control, it is possible that GFP is not expressed or the N-terminus peptides interrupt its function. The expression of GFP can be tested by western blotting with GFP antibody. If GFP is expressed but not functional, longer linker will be used to ensure that the artificial peptides do not interrupt GFP function. Alternatively, a self-cleaving picornavirus T2A peptide which cleaves itself during translation (Griffioen, M., et al. Genetic engineering of virus-specific T cells with T-cell receptors recognizing minor histocompatibility antigens for clinical application. *Haematologica* 93, 1535-1543 (2008)), can be used as the linker. An additional method is to generate a new zinc finger that recognizes 18 base pairs of sequence in the beginning of gfp, which might avoid the peptide interruption problem.

[0107] Synthetic genes encoding *Escherichia coli* codon-optimized humanized AID (hAID) and OCT_ZFP will be generated (DNA 2.0 inc.). A variety of aid-zfp with different lengths of linkers (G3S)n in the coding region will be constructed by overlap-extension PCR. These constructs will be cloned into pET-DEST42 and transformed into the bacteria generated (described above). For simplicity, UNG inhibitor will also be expressed to inhibit the repair pathway (Petersen-Mahrt, S. K., Harris, R. S. & Neuberger, M. S. AID mutates *E. coli* suggesting a DNA deamination mechanism for antibody diversification. *Nature* 418, 99-103 (2002)). The transformation will be verified by antibiotic resistant selection. Florescent microscopy and flow cytometry will be applied to test whether GFP signal is rescued. Monocolonies with GFP^+ signal will be sorted out followed by sequence analysis to verify whether the rescue of GFP is the result of reversing the mutated start codon from "ACG" to "ATG" (FIG. **2**). As a negative control, ZF will be expressed alone to assess the rate of spontaneous mutation. For comparison, AID will be expressed alone in the cell to assess whether the sequence context of start codon introduces bias.

[0108] Without intending to be bound by scientific theory, if AID-ZFP^{oct4} can introduce more GFP^+ cells than ZFP alone, this will indicate that AID is active in the fusion protein. To determine the targeted mutation efficacy, the GFP^+ cell percentage under the expression of AID alone will be taken into consideration in the analysis. As shown in Table 1, A, B, C and D represent the GFP rescue efficiency under different conditions.

TABLE 1

the percentage of GFP^+ cell under different conditions

	protein	
Genotype of GFP	AID-ZFP	AID
Zinc finger recognition site* Broken start codon	A	B
Altered ZF recognition site* Broken start codon	C	D

[0109] When gfp start codon is in a random sequence context, the GFP rescue efficiency (C and D) represents the deamination activity. When gfp start codon is in the zinc finger targeting site, both the deamination and targeting efficacy contribute to the to the GFP rescue efficiency (A and B). As a result, the efficacy of AID-ZF targeting can be resolved as: Efficacy (E)=(A/B)/(C/D). If E>1, AID-ZFP can specifically target the cytidine. If E<1, there is no targeted mutation. An alternative approach to analyze the targeting efficacy of AID-ZFP is to construct and express AID-NZF, in which ZFP loses its DNA binding (Green, A. & Sarkar, B. Alteration of zif268 zinc-finger motifs gives rise to non-native zinc-coordination sites but preserves wild-type DNA recognition. *Biochem J* 333 (Pt 1), 85-90 (1998)). The direct comparison of GFP rescue efficiency between AID-ZFP and AID-NZF will decipher the targeting efficacy of AID-ZFP. However, the presumption of this design is that both AID-ZFP and AID-NZF have similar deamination activity, which is not necessarily true.

[0110] Without intending to be bound by scientific theory, there are many factors that may potentially contribute to this result. (1) It is possible that zinc finger cannot find its right target in vivo. To test the first possibility, chromatin immunoprecipitation (ChiP) experiment will be performed. If ChiP indicates AID-ZFP cannot bind to its target site, different lengths of linker will be tested to find a proper structure in which AID and ZF do not interrupt each other's function. (2) It is possible that AID loses the deamination activity. Longer linker will be used if AID activity is the problem. Alternatively, AID can be fused to the C-terminus of ZFP if AID cannot function properly in the N-terminus. (3) It is possible that AID functions as a dimer, thus the recruitment of a single copy of AID-ZF is not sufficient to trigger significant deamination reaction. If this is the case, an artificial dimer will be generated by building an AID-AID-ZFP construct. Alternatively, two different zinc fingers can be designed to bind the upstream and downstream of the target site. The binding of the two different AID-ZFPs to this region will force AID to dimerize in the middle and deaminate the targeted cytidine.

[0111] In certain aspects, APOBEC1 will be used instead of AID. Although APOBEC1 was thought to be a RNA deaminase, recent studies show that APOBEC1 can deaminate cytidine in DNA in vitro (Petersen-Mahrt, S. K. & Neuberger, M. S. In vitro deamination of cytosine to uracil in single-stranded DNA by apolipoprotein B editing complex catalytic subunit 1 (APOBEC1).

Example IV

Testing Whether AID-ZFP can Specifically Target Sites in the Bacterial Genome Without Introducing Toxic Effect(s)

[0112] ChIP sequencing will be performed to identify all locations in the genome to which the AID-ZFP binds. Briefly, AID-ZFP will be tagged with His on its C-terminus and be cloned into pET-DEST42. AID-ZFP-HIS will be expressed in the bacterial system that is constructed as described above. Subsequently, tagged AID-ZFP will be cross-linked to the bound DNA in vivo, the cell will be lysed, and the DNA be sheared. Later, anti-His antibodies will be used to pull down the protein-DNA complex. The identities of bound DNA and the percent occupancy of the AID-ZFN at these locations will be resolved by sequencing. For comparison, tagged ZFP will be conducted in parallel.

[0113] If AID does not interfere with the binding between ZFP and DNA, AID-ZFP will exhibit the similar binding pattern as ZFP. If AID-ZFP shows less affinity to ZFP binding site and increased off-target rate, it indicates AID interferes the DNA binding ability. Without intending to be bound by scientific theory, it is possible that AID and ZFP are too close, thus each module cannot function properly. In this case, longer linker can be tested. Also, the structure of AID might distort the binding specificity. Without intending to be bound by scientific theory, it is possible that the chemical property of AID N-terminus is responsible for the distortion of AID-ZFP DNA binding specificity. Proper engineering of the N-terminus of AID can reduce its tendency to bind to DNA, thus reduce the interruption.

[0114] Protein binding microarray (PBM) assays can be used to systematically test AID-ZFP binding specificity in vitro. Essentially hAID-ZFP-HIS will be expressed and purified. A dsDNA microarray that has several thousand dsDNA variants (the target sequence+all 54 one position variants [54=18*3]$^{+}$ all 1377 two position variants [1377=18*17/2*9], all 14688 three position variants=16120) will be generate. The array will then be incubated with AID-ZFP-HIS and Cy3-conjugated mouse anti-His monoclonal antibody (Sigma) subsequently. The binding affinity of AID-ZFP to different sequence can be measured by the florescent density of each dot on the array.

[0115] AID-ZFP with different linkers ((G3S)n, LRGS, $G(SGGGG)_2$, SGGGLGST and the like) will be constructed individually and expressed in bacteria that has a GFP reporter system. Monocolonies of GFP^{+} cells will be selected and the gfp gene will be sequenced to test whether the cytidine residues near the targeting site are deaminated.

[0116] Without intending to be bound by scientific theory, AID-ZFP constructs with shorter linkers are expected to have less or even no wobble targeting, because there is less room for AID to slide along the ssDNA. Further shortening of the linker will compromise the deamination activity if the two modules are too close together to function properly. If AID still deaminates the neighboring cytidines regardless of the length of the linker, the AID mutants R35E and R35E/R36D that have less processivity (Bransteitter, R., Pham, P., Calabrese, P. & Goodman, M. F. Biochemical analysis of hypermutational targeting by wild type and mutant activation-induced cytidine deaminase. *J Biol Chem* 279, 51612-51621 (2004)) will be generated and tested. An alternative method to look for evidence of progressive AID events is to look for, count, and analyze different sectors of sectored colonies.

[0117] The GFP reporter system described herein will be utilized for the expression of AID-ZFP and the negative control ZFP will be driven by PL-tetO promoter, which can modulate gene expression with a linear response when paired with tetR-aTc protein-small molecule (Lutz, R. & Bujard, H. Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I142 regulatory elements. *Nucleic Acids Res* 25, 1203-1210 (1997)). The expression level of AID-ZFP will be assayed by QPCR. Cell growth rate will be measured by spectrometry and GFP^{+} cell percentage will be measured by flow cytometry. Subsequently, the genomic DNA of mono-colony GFP^{+} cell with different expression level of AID-ZFP will be extracted and sheared. Size selected DNA fragments will be ligated with barcoded adaptors and the whole genome sequencing will be performed to analyze the off-target mutagenesis profile.

[0118] Toxic ZFNs reduce both the percentage of GFP-positive cells and cells that have undergone gene targeting. The toxicity of AID-ZFP can also be measured by the viability of GFP^+ cells and growth rate of cells that are transformed. As FIG. **3**, with the increase of AID-ZF expression level, the GFP^+ cell percentage will increase, but cell toxicity will also increase so that GFP^+ cell percentage will arrive at a plateau or even drop back. Optimized AID-ZFP expression level will be selected and further analyzed by sequencing. One illumine sequencing reaction generates[53] 2,160,000,000 by reads, which covers the bacteria genome 480 times (2160/4.5=480). Assuming that 10 times coverage is sufficient to place a read in the genome, 48 different *E. coli* strains can be sequenced. Comparison the genome sequence in the AID-ZFP expressed strain (different levels) with that of ZFP expressed strain will reveal the off-target mutation rate. One pitfall with this experiment is that the sequence error and the heterogeneity of different bacteria will introduce false positive. A complementary method is to perform ChIP-seq using a version of epitope-labeled UNG that lacks activity. This UNG would specifically bind to uracils and pull down the uracil containing fragments of the DNA, which will then be sequenced and located.

[0119] Whether or not AID-ZFP can be used as a targeting mutator to introduce a specific C to T mutation in K-ras, a gene which is mutationally activated in approximately 20% of all solid tumors, will be determined. Aberrant activation of K-ras signaling pathway has been strongly implicated in the pathogenesis of neoplasia in the lung, pancreases, and colon. However, the development of clinically effective K-RAS-directed cancer therapies has been largely unsuccessful and K-ras mutant cancers remain among the most refractory to available treatment. AID-ZFP can be used in mammalian cells to specifically introduce a premature stop codon (UAG to TAG mutation) in K-ras gene and abolish its function. First, the targeting efficacy by a GFP assay in HEK 293 cells will be assessed. Next, the specificity of AID-ZFP targeting will be examined. Finally, whether AID-ZFP can abolish the translation of K-Ras, thus inhibiting cell growth, will be ascertained.

Example V

Determining Whether AID-ZFP can Change a Specific Cytidine Residue in the K-ras120-egfp in HEK 293 Cells

[0120] First, 293-K-ras120-egfp cells with K-ras120-egfp gene incorporated into the HEK 293 genome will be generated (FIG. **4**A). Essentially, the first 120 base pairs of K-ras protein coding region (120 out of 566) will be fused with egfp using a linker in between. This construct will be cloned into the pcDNA5/FRT/TO vector, which will then be co-transfected into HEK 239 Flp-In cell. Cells that incorporate K-ras120-egfp will be selected by GFP signal and hygromycin resistance. As a control, $K\text{-}ras120^{stop}$-egfp will be constructed in parallel in which the Gln22 (CAG) is mutated to a stop codon (TAG) to ensure that the introduction of a premature stop codon abolishes the translation of EGFP.

[0121] Second, AID-$ZFP^{K\text{-}ras}$ construct will be created (FIG. **4**B). Briefly, 6 zinc finger arrays that target the Gln22 coding region (CAG) of K-ras gene will be assembled. $ZFP^{K\text{-}ras}$ will be fused with AID by linkers of various lengths. Subsequently, AID-$ZFP^{K\text{-}ras}$ will be cloned into pMSCV-IRES-RFP-puro vector (Holliday, R. & Grigg, G. W. DNA methylation and mutation. *Mutat Res* 285, 61-67 (1993)) and then delivered into the 293-K-ras120-egfp cells. Transfected cell will be selected by $puro^R$ and flow cytometry will be performed to measure yellow cell (GFP RFP) (FIG. **4**B) and red cell (RFP) (FIG. **4**C) percentages. To verify the targeted mutation on K-ras gene, the first 120 base pairs in the K-ras120 CDS region (both in the K-ras120-egfp, and endogenous K-ras120 gene) will be sequenced. As a negative control, $ZFP^{K\text{-}ras}$ alone will be expressed in parallel with AID-$ZFP^{K\text{-}ras}$ to evaluate the rate of mutations introduced by factors other than AID. For comparison, a parallel construct of AID-NZF in which ZFP cannot bind to any DNA sequence will be expressed to examine the target efficiency of AID-$ZFP^{K\text{-}ras}$.

[0122] RFP^+ cells represent the cells in which AID-$ZFP^{K\text{-}ras}$ successfully mutated the GFP gene (FIG. **5**B). If the RFP cell percentage is higher in the AID-$ZFP^{K\text{-}ras}$ group than that of the $ZFP^{K\text{-}ras}$ group, it indicates that AID is active in the fusion construct. If the RFP cell percentage is higher in the $ZFP^{K\text{-}ras}$ group than that of the AID-NZF group, it indicates $ZFP^{K\text{-}ras}$ helps AID to specifically target the K-ras gene. Sequence analysis will further verify whether the loss of GFP signal is a result of CAG to TAG transition in the Gln22 position on K-ras120-egfp gene. Successful targeting should also result in another CAG to TAG mutation in the endogenous K-ras genes.

[0123] Without intending to be bound by scientific theory, if the RFP cell percentage is the same or even lower in the AID-$ZFP^{K\text{-}ras}$ expression group than that in the AID-NZF group, it suggests that AID-ZFP cannot target the K-ras-egfp gene. Besides the possible reasons discussed above, there are some special factors in the human cell system that might account for this result. (1) AID-ZFP cannot get into the nucleus. Since AID harbors a natural nucleus localization signal (NLS) at its N-terminus, AID-ZFP should be transported into the nucleus. However, it is possible that in the fusion protein, the NLS cannot interact with the nucleus transportation factors properly due to the interference of ZFP, thus failing to enter the nucleus. To test this possibility, AID-ZFP tagged with a V5 epitope will be expressed and its location will be visualized by incubating the cells with fluorescence-labeled V5 antibody. If the localization of AID-ZFP is a problem, the artificial NLS that is used in the ZFN system will be applied to the AID-ZFP construct to enhance the transportation signal. (2) Cellular repair systems, such as base excision repair (BER) or mismatch repair (MMR) pathways might recognize the uridine introduced by AID-ZFP, and repair it before it can be resolved to thymidine. To test this possibility, UNG and MSH2 will be transiently knocked down by siRNA separately to test whether these repair machineries fix the mutations introduced by AID. (3) Chromosomal structure or target site methylation would affect the accessibility of the target sites to AID-ZFP. To test this possibility, a ChiP-Seq experiment (as discussed further herein) will be performed to assess the DNA binding situation of AID-$ZFP^{K\text{-}ras}$. Under these circumstances, multiple sites on K-ras will be chosen to target in case an AID-ZFP cannot bind to its targeting site due to the local chromosome structure. Comparing the effect of AID-$ZFP^{K\text{-}ras}$ and AID-NZP will enable the distinction between the targeted mutation effect and the random mutation effect. The EGFP gene will be sequenced to determine the causative mutations.

[0124] The specificity of AID-ZFP will be determined by examining the off-target rates. AID-$ZFP^{K\text{-}ras}$ will be tagged with HIS and be expressed in the 293-K-ras120-egfp cells. RFP^+ cell will be isolated and cultured. AID-$ZFP^{K\text{-}ras}$ will be cross-linked with the DNA that it binds to and the DNA-protein complex will be pulled down by anti-His antibody.

Deep sequencing will reveal the binding sites of AID-ZFP$^{K\text{-}ras}$ throughout the genome. As a positive control, ZFP$^{K\text{-}ras}$ will be processed in parallel with AID-ZFP$^{K\text{-}ras}$

[0125] Without intending to be bound by scientific theory, if AID-ZFP$^{K\text{-}ras}$ retains the DNA binding specificity of ZFP$^{K\text{-}ras}$, the majority of bound DNAs should represent the ZFP target sites. Moreover, the off-target sites are likely to be the sites that share similar sequences with ZFP$^{K\text{-}ras}$ recognition sites. One caveat about this experiment is that it only measures the ZFP$^{K\text{-}ras}$ binding specificity not the deamination specificity of this whole construct. It is possible that AID deaminates random sites regardless whether it has strong binding affinity to those sites. For example, AID might interact with certain factors that recruit it to some positions in the genome while the interaction of AID-cofactor-DNA is not strong enough to be revealed by this ChiP-seq. In addition, since it only measures the binding specificity, the sequences that pulled down are not necessarily deaminated.

[0126] An alternative way to analyze the specificity of AID-ZFP is genome-wide Chip-seq using a version of epitope-labeled uracil-DNA glycosylases (UNG) that lacks activity. This UNG would bind to uracils and pull down AID modified fragments of the DNA that could then be sequenced and located. This ChiP-seq will reveal the deamination specificity of AID-ZFP$^{K\text{-}ras}$.

[0127] The capan-1 cell line is a pancreatic tumor cell line that has aberrant activation of K-RAS will be transfected with pMSCV-AID-ZFP$^{K\text{-}ras}$-IRES-RFP-euro, and the transfected cells will be selected by RFP^{+} and Puro^{r}. As a negative control, cells will be transfected with pMSCV-AID-NZP-IRES-RFP-puro in parallel. As a positive control, cells will be infected with lentiviral particles that have shRNA$^{k\text{-}ras}$. To determine whether AID-ZFP$^{K\text{-}ras}$ introduces a premature stop codon into the K-ras gene, the K-ras gene will be sequenced and the mRNA levels of K-ras will be measured by quantitative PCR (QPCR). To determine whether the premature stop codon abolishes K-RAS function, the protein level and size of K-RAS will be tested by western. To determine whether mutated K-RAS inhibits the growth of cells, cell proliferation will be assayed in triplicate using Brdu-cytometry, and cell apoptosis will be measured by Casp-3 signaling by flow cytometry.

[0128] If AID-ZFP$^{K\text{-}ras}$ can specifically target K-ras, the transition of CAG to TAG will be observed in the K-ras gene. Also, the mRNA expression level of K-ras is supposed to decrease due to nonsense mediated decay (NMD). In the Western experiment, the truncated K-RAS protein (2.2 KD) should be detected instead of the full length K-RAS (21 KD). If this truncated K-RAS loses function, cell growth rate will decrease, while the apoptosis signal will increase.

[0129] If K-Ras does not lose its activity, another AID-ZFP$^{K\text{-}ras\text{-}start}$ construct will be built to mutate the start codon from ATG to ATA. If the introduction of AID-ZFP$^{K\text{-}ras}$ inhibits cell growth and triggers apoptosis, rescue experiments will be conducted to test the targeting specificity and toxicity of AID-ZFP$^{k\text{-}ras}$. Another copy of K-Ras cDNA, which has silent mutations that lose the binding site of AID-ZFP$^{k\text{-}ras}$ will be introduced into the cell (FIG. **5**). If AID-ZFP$^{k\text{-}ras}$ specifically targets the endogenous K-Ras and has no other undefined toxic effects, the exogenous K-Ras will rescue the cell so that the cell growth rate and apoptosis signaling will go back to normal level.

REFERENCES

[0130] Delker, R. K., Fugmann, S. D. & Papavasiliou, F. N. A coming-of-age story: activation-induced cytidine deaminase turns 10. *Nat Immunol* 10, 1147-1153 (2009).

[0131] Muramatsu, M., et al. Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells. *J Biol Chem* 274, 18470-18476 (1999).

[0132] Stavnezer, J., Guikema, J. E. & Schrader, C. E. Mechanism and regulation of class switch recombination. *Annu Rev Immunol* 26, 261-292 (2008).

[0133] Storb, U., et al. Targeting of AID to immunoglobulin genes. *Adv Exp Med Biol* 596, 83-91 (2007).

[0134] Teng, G. & Papavasiliou, F. N. Immunoglobulin somatic hypermutation. *Annu Rev Genet.* 41, 107-120 (2007).

[0135] Bransteitter, R., Pham, P., Scharff, M. D. & Goodman, M. F. Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. *Proc Natl Acad Sci USA* 100, 4102-4107 (2003).

[0136] Ramiro, A. R., Stavropoulos, P., Jankovic, M. & Nussenzweig, M. C. Transcription enhances AID-mediated cytidine deamination by exposing single-stranded DNA on the nontemplate strand. *Nat Immunol* 4, 452-456 (2003).

[0137] Shen, H. M. & Storb, U. Activation-induced cytidine deaminase (AID) can target both DNA strands when the DNA is supercoiled. *Proc Natl Acad Sci USA* 101, 12997-13002 (2004).

[0138] Peled, J. U., et al. The biochemistry of somatic hypermutation. *Annu Rev Immunol* 26, 481-511 (2008).

[0139] Yoshikawa, K., et al. AID enzyme-induced hypermutation in an actively transcribed gene in fibroblasts. *Science* 296, 2033-2036 (2002).

[0140] Jovanic, T., Roche, B., Attal-Bonnefoy, G., Leclercq, 0. & Rougeon, F. Ectopic expression of AID in a non-B cell line triggers A:T and G:C point mutations in non-replicating episomal vectors. *PLoS One* 3, e1480 (2008).

[0141] Klasen, M., Spillmann, F. J., Marra, G., Cejka, P. & Wabl, M. Somatic hypermutation and mismatch repair in non-B cells. *Eur J Immunol* 35, 2222-2229 (2005).

[0142] Martin, A. & Scharff, M. D. Somatic hypermutation of the AID transgene in B and non-B cells. *Proc Natl Acad Sci USA* 99, 12304-12308 (2002).

[0143] Cathomen, T. & Joung, J. K. Zinc-finger nucleases: the next generation emerges. *Mol Ther* 16, 1200-1207 (2008).

[0144] Lee, M. S., Mortishire-Smith, R. J. & Wright, P. E. The zinc finger motif. Conservation of chemical shifts and correlation with structure. *FEBS Lett* 309, 29-32 (1992).

[0145] Jayakanthan, M., et al. ZifBASE: a database of zinc finger proteins and associated resources. *BMC Genomics* 10, 421 (2009).

[0146] Pabo, C. O., Peisach, E. & Grant, R. A. Design and selection of novel Cys2His2 zinc finger proteins. *Annu Rev Biochem* 70, 313-340 (2001).

[0147] Foley, J. E., et al. Rapid mutation of endogenous zebrafish genes using zinc finger nucleases made by Oligomerized Pool ENgineering (OPEN). *PLoS One* 4, e4348 (2009).

[0148] Moehle, E. A., et al. Targeted gene addition into a specified location in the human genome using designed zinc finger nucleases. Proc Natl Acad Sci USA 104, 3055-3060 (2007).

[0149] Maeder, M. L., Thibodeau-Beganny, S., Sander, J. D., Voytas, D. F. & Joung, J. K. Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays. *Nat Protoc* 4, 1471-1501 (2009).

[0150] Hockemeyer, D., et al. Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nat Biotechnol* 27, 851-857 (2009).

[0151] Wright, D. A., et al. High-frequency homologous recombination in plants mediated by zinc-finger nucleases. *Plant J*44, 693-705 (2005).

[0152] Maeder, M. L., et al. Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol Cell* 31, 294-301 (2008).

[0153] Xu, G. L. & Bestor, T. H. Cytosine methylation targetted to pre-determined sequences. *Nat Genet.* 17, 376-378 (1997).

[0154] Harper, J., et al. Repression of vascular endothelial growth factor expression by the zinc finger transcription factor ZNF24. *Cancer Res* 67, 8736-8741 (2007).

[0155] Dhanasekaran, M., Negi, S. & Sugiura, Y. Designer zinc finger proteins: tools for creating artificial DNA-binding functional proteins. *Acc Chem Res* 39, 45-52 (2006).

[0156] Kim, H. J., Lee, H. J., Kim, H., Cho, S. W. & Kim, J. S. Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly. *Genome Res* 19, 1279-1288 (2009).

[0157] Zeng, X. & Rao, M. S. Controlled genetic modification of stem cells for developing drug discovery tools and novel therapeutic applications. *Curr Opin Mol Ther* 10, 207-213 (2008).

[0158] Wang, H. H., et al. Programming cells by multiplex genome engineering and accelerated evolution. *Nature* 460, 894-898 (2009).

[0159] Fajardo-Sanchez, E., Stricher, F., Paques, F., Isalan, M. & Serrano, L. Computer design of obligate heterodimer meganucleases allows efficient cutting of custom DNA sequences. *Nucleic Acids Res* 36, 2163-2173 (2008).

[0160] Groth, A. C. & Calos, M. P. Phage integrases: biology and applications. *J Mol Biol* 335, 667-678 (2004).

[0161] Kandavelou, K., et al. Targeted manipulation of mammalian genomes using designed zinc finger nucleases. Biochem *Biophys Res Commun* 388, 56-61 (2009).

[0162] Yanez, R. J. & Porter, A. C. Therapeutic gene targeting. *Gene Ther* 5, 149-159 (1998).

[0163] Li, J. B., et al. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* 324, 1210-1213 (2009).

[0164] Luo, J., Solimini, N. L. & Elledge, S. J. Principles of cancer therapy: oncogene and non-oncogene addiction. *Cell* 136, 823-837 (2009).

[0165] Holliday, R. & Grigg, G. W. DNA methylation and mutation. *Mutat Res* 285, 61-67 (1993).

[0166] Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet.* 10, 691-703 (2009).

[0167] Rada, C., Jarvis, J. M. & Milstein, C. AID-GFP chimeric protein increases hypermutation of Ig genes with no evidence of nuclear localization. *Proc Natl Acad Sci USA* 99, 7003-7008 (2002).

[0168] Griffioen, M., et al. Genetic engineering of virus-specific T cells with T-cell receptors recognizing minor histocompatibility antigens for clinical application. *Haematologica* 93, 1535-1543 (2008).

[0169] Petersen-Mahrt, S. K., Harris, R. S. & Neuberger, M. S. AID mutates *E. coli* suggesting a DNA deamination mechanism for antibody diversification. *Nature* 418, 99-103 (2002).

[0170] Green, A. & Sarkar, B. Alteration of zif268 zinc-finger motifs gives rise to non-native zinc-co-ordination sites but preserves wild-type DNA recognition. *Biochem J* 333 (Pt 1), 85-90 (1998).

[0171] Petersen-Mahrt, S. K. & Neuberger, M. S. In vitro deamination of cytosine to uracil in single-stranded DNA by apolipoprotein B editing complex catalytic subunit 1 (APOBEC1). *J Biol Chem* 278, 19583-19586 (2003).

[0172] Harris, R. S., Petersen-Mahrt, S. K. & Neuberger, M. S. RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. *Mol Cell* 10, 1247-1253 (2002).

[0173] Teng, B. B., et al. Mutational analysis of apolipoprotein B mRNA editing enzyme (APOBEC1). structure-function relationships of RNA editing and dimerization. *J Lipid Res* 40, 623-635 (1999).

[0174] Storb, U., Shen, H. M. & Nicolae, D. Somatic hypermutation: processivity of the cytosine deaminase AID and error-free repair of the resulting uracils. *Cell Cycle* 8, 3097-3101 (2009).

[0175] Chelico, L., Pham, P. & Goodman, M. F. Stochastic properties of processive cytidine DNA deaminases AID and APOBEC3G. *Philos Trans R Soc Lond B Biol Sci* 364, 583-593 (2009).

[0176] Pham, P., Bransteitter, R., Petruska, J. & Goodman, M. F. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* 424, 103-107 (2003).

[0177] Bransteitter, R., Pham, P., Calabrese, P. & Goodman, M. F. Biochemical analysis of hypermutational targeting by wild type and mutant activation-induced cytidine deaminase. *J Biol Chem* 279, 51612-51621 (2004).

[0178] Lutz, R. & Bujard, H. Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I142 regulatory elements. *Nucleic Acids Res* 25, 1203-1210 (1997).

[0179] Cornu, T. I., et al. DNA-binding specificity is a major determinant of the activity and toxicity of zinc-finger nucleases. *Mol Ther* 16, 352-358 (2008).

[0180] Pruett-Miller, S. M., Connelly, J. P., Maeder, M. L., Joung, J. K. & Porteus, M. H. Comparison of zinc finger nucleases for use in gene targeting in mammalian cells. *Mol Ther* 16, 707-717 (2008).

[0181] Handel, E. M., Alwin, S. & Cathomen, T. Expanding or restricting the target site repertoire of zinc-finger nucleases: the inter-domain linker as a major determinant of target site selectivity. *Mol Ther* 17, 104-111 (2009).

[0182] Fan, J. B., Chee, M. S. & Gunderson, K. L. Highly parallel genomic assays. *Nat Rev Genet.* 7, 632-644 (2006).

[0183] Wardle, J., et al. Uracil recognition by replicative DNA polymerases is limited to the archaea, not occurring with bacteria and eukarya. *Nucleic Acids Res* 36, 705-711 (2008).

[0184] Singh, A., et al. A gene expression signature associated with "K-Ras addiction" reveals regulators of EMT and tumor cell survival. *Cancer Cell* 15, 489-500 (2009).

SEQUENCE LISTING

<160> NUMBER OF SEQ ID NOS: 2

<210> SEQ ID NO 1
<211> LENGTH: 84
<212> TYPE: PRT
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: AID fusion protein
<220> FEATURE:
<221> NAME/KEY: MISC_FEATURE
<222> LOCATION: (17)..(23)
<223> OTHER INFORMATION: wherein X is any amino acid
<220> FEATURE:
<221> NAME/KEY: MISC_FEATURE
<222> LOCATION: (45)..(51)
<223> OTHER INFORMATION: wherein X is any amino acid
<220> FEATURE:
<221> NAME/KEY: MISC_FEATURE
<222> LOCATION: (73)..(79)
<223> OTHER INFORMATION: wherein X is any amino acid

<400> SEQUENCE: 1

Pro Gly Glu Arg Pro Phe Gln Cys Arg Ile Cys Met Arg Asn Phe Ser
1 5 10 15

Xaa Xaa Xaa Xaa Xaa Xaa Xaa His Thr Arg Thr His Thr Gly Glu Lys
20 25 30

Pro Phe Gln Cys Arg Ile Cys Met Arg Asn Phe Ser Xaa Xaa Xaa Xaa
35 40 45

Xaa Xaa Xaa His Leu Arg Thr His Thr Gly Glu Lys Pro Phe Gln Cys
50 55 60

Arg Ile Cys Met Arg Asn Phe Ser Xaa Xaa Xaa Xaa Xaa Xaa Xaa His
65 70 75 80

Leu Lys Thr His

<210> SEQ ID NO 2
<211> LENGTH: 18
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: OCT4 ZFP binding site

<400> SEQUENCE: 2

gagcaggcag ggtcagct 18

What is claimed is:

1. A non-naturally occurring fusion protein comprising:

a DNA binding domain; and

a DNA modifying domain that includes a functional fragment of a deaminase protein, wherein the fusion protein is capable of binding to and altering a target oligonucleotide sequence.

2. The fusion protein of claim **1**, wherein the DNA binding domain includes a motif selected from the group consisting of helix-turn-helix, leucine zipper, winged helix, winged helix turn helix, helix-loop-helix, zinc finger, immunoglobulin fold, B3 domain and TATA-box binding protein domain.

3. The fusion protein of claim **1**, wherein the deaminase protein is activation-induced deaminase (AID).

4. The fusion protein of claim **1**, wherein the target oligonucleotide sequence is DNA.

5. The fusion protein of claim **4**, wherein the DNA is genomic DNA.

6. An isolated polynucleotide encoding the fusion protein of claim **1**.

7. An expression vector comprising the isolated polynucleotide of claim **6**.

8. A host cell expressing the expression vector of claim **7**.

9. A cell comprising a non-naturally occurring fusion protein, wherein the fusion protein includes a DNA binding domain, and a DNA modifying domain that includes a functional fragment of a deaminase protein, wherein the fusion protein is capable of binding to and altering a target oligonucleotide sequence.

10. The cell of claim **9**, wherein the deaminase protein is AID.

11. The cell of claim **9**, wherein the cell is an animal cell.

12. The cell of claim **11**, wherein the animal cell is a mammalian cell.

13. The cell of claim **12**, wherein the mammalian cell is a human cell.

14. The cell of claim **9**, wherein the cell is a stem cell.

15. The cell of claim **14**, wherein the stem cell is a hematopoietic stem cell.

16. A method of modulating expression of an endogenous gene in a cell, comprising the steps of:

contacting a cell with a non-naturally occurring fusion protein wherein the fusion protein includes a DNA binding domain, and a DNA modifying domain including a functional fragment of a deaminase protein, wherein the fusion protein is capable of binding to and altering an oligonucleotide sequence of an endogenous gene; and

allowing the fusion protein to bind to and alter the oligonucleotide sequence of the endogenous gene to modulate expression of the endogenous gene.

17. The method of claim **16**, wherein the deaminase protein is AID.

18. The method of claim **16**, wherein the cell is an animal cell.

19. The method of claim **18**, wherein the animal cell is a mammalian cell.

20. The method of claim **19**, wherein the mammalian cell is a human cell.

21. The method of claim **16**, wherein the cell is a stem cell.

22. The method of claim **21**, wherein the stem cell is a hematopoietic stem cell.

23. The method of claim **16**, wherein expression of the endogenous gene is repressed.

24. The method of claim **16**, wherein expression of the endogenous gene is activated.

* * * * *